An automated tiered-evidence toolkit for cybersecurity compliance assessment in Saudi SMEs

AlJasser AlGhamdi

Abstract

Small and medium-sized enterprises (SMEs) in Saudi Arabia must demonstrate compliance with the Essential Cybersecurity Controls (ECC-2:2024) and the companion standard for non-critical private-sector entities (NCNICC-1:2025). Most operate without dedicated security personnel. Unauthenticated external scanning provides relatively low-cost independent evidence. We examine which controls in the two standards it can test. We classify each control row by the evidence required to assess it: external observation, internal assessment, or a signed declaration. External observation can test 16 of the 92 ECC-2:2024 subcontrols. When the seven parent controls are included, this corresponds to 23 of the 203 catalogue rows (11.3%). The remaining 180 rows require internal artefacts or declarations. We release three supporting artefacts: a mapping indexed by control identifier, an automated toolkit that produces per-control evidence, and a synthetic cohort of 100 Saudi Class B firms based on national demographic statistics. On this synthetic cohort the toolkit achieves a modelled micro-F_1 of 0.909 for detecting violations of the externally testable controls, compared with 0.732 for a predict-all baseline that labels every control as violated. Among declarations concerning controls the scan can observe, it flags contradictions at F_1 0.898. Both results are end-to-end integrity checks under a parameterised sensing model, not field measurements; 15 of the 31 detectors are implemented. The small externally testable share is a central finding: it shows where external observation ends and where organisation-supplied evidence must begin. The classification is author-curated and not yet validated by an independent expert panel.

Keywords: compliance evidence, external attack surface, automated security assessment, evidence tiering, SME cybersecurity, ECC-2:2024, NCNICC-1:2025, Saudi Arabia

1. Introduction

Demonstrating cybersecurity compliance requires auditable evidence for each control. Standards and frameworks such as ISO/IEC 27001:2022, the NIST Cybersecurity Framework 2.0 and the CIS Critical Security Controls define the expected practices, but implementation alone does not demonstrate compliance: each control must be supported by evidence an assessor can audit [1–3]. Producing this evidence is manual, repeated and dependent on specialist expertise, and the burden grows with the number of controls rather than with the organisation's size. Researchers have therefore treated control text as structured data and evidence collection as an automatable workflow [4–8].

This gap between regulatory obligations and organisational capacity is greatest among small and medium-sized enterprises (SMEs), which must meet the same control sets as large organisations but have far fewer security resources. Studies across several jurisdictions show the same pattern: SMEs report acute shortages of staff, tooling and process [9–11], they under-invest relative to their exposure [12], and expert assessment shows that their reported posture differs systematically from observed conditions [13]. This last finding matters for any assessment tool: most compliance tooling relies on the organisational self-report, which carries a documented upward bias.

Three assessment approaches address different parts of this evidence problem.

The first asks respondents to report the status of each control and scores the responses. Risk-weighted compliance assessment systems, maturity models and questionnaire-driven self-assessment tools follow this design across national and sectoral frameworks [14–18], and some recent variants use a single self-declared maturity value to recommend controls [19, 20]. These tools are inexpensive to deploy and require no technical access or setup. Their output, however, depends on the accuracy of the self-report: the assessment is only as reliable as the respondent's account of the estate (the organisation's systems and internet-facing assets). An independent check is therefore needed before the result can serve as evidence rather than as a claim.

The second collects evidence from inside the administrative boundary and evaluates it against machine-readable control representations. Structured formats such as OSCAL and SCAP allow software to process individual controls [4, 21, 22]; knowledge-based methods use those representations to check policy against configuration [23]; and continuous compliance systems evaluate cloud controls against live telemetry on a schedule [5, 6, 8]. These systems automate evidence collection at control granularity. They do so by operating with privileged access: an agent, an API credential, a configuration export or a telemetry pipeline. They also need an operator competent

to deploy and maintain them. The requirements are therefore organisational as well as technical, and many SMEs cannot meet them.

The third observes organisations from outside, requiring neither cooperation nor credentials. Internet-scale measurement enumerates reachable services and reads their configuration directly [24–27]; attack-surface management focuses on the unknown and unmanaged internet-facing assets an organisation holds [28–31]; and externally visible security weaknesses have repeatedly been shown to predict organisational risk and breach likelihood [32–36]. Surveys describe both the capabilities and the methodological limitations of these techniques [37, 38]. These methods produce finding lists or composite risk scores based on observed exposures: a stale certificate, an absent authentication record, or a reachable management port. Commercial security-rating services compress these findings into a single score whose derivation is opaque and whose values diverge across providers [39]. Converting such observations into compliance evidence therefore requires a further step: a mapping from observed exposures to the controls an assessment regime uses. Where that mapping has been attempted, it has been at the level of an overall certification verdict rather than individual controls, as work on payment-card certification and on national assurance schemes illustrates [40, 41].

We have not identified an existing approach that combines external observation with evidence at control granularity. Such a tool must satisfy four requirements. It must operate without privileged access, so that organisations without security staff can use it. It must produce evidence linked to control identifiers rather than to exposures, so that formal assessment processes can use the output. It must be explicit about which controls it can test and which it cannot, since external observation provides no evidence about governance, personnel, physical or backup controls, whatever the quality of the measurement. And it must incorporate organisation-supplied evidence for the controls it cannot test, rather than leaving them unaddressed. Building such a tool first requires a quantity that, to our knowledge, has not been established for any national control set: the proportion of controls that unauthenticated external observation can test at all.

This work examines the Saudi regulatory regime, where the gap between obligation and capacity is acute and the two control sets are small enough for exhaustive analysis. The National Cybersecurity Authority (NCA) issued the second revision of the Essential Cybersecurity Controls (ECC-2:2024) together with a companion standard for non-critical private-sector entities (NCNICC-1:2025) [42, 43]. The NCNICC-1:2025 Class B tier covers organisations employing 6 to 249 staff or reporting annual revenue between SAR 3 million and SAR 200 million. ECC-2:2024 is organised into 4 domains and 28 subdomains. It contains 108 main controls and 92 subcontrols, for 200 control rows. NCNICC-1:2025 marks each control mandatory or recommended per class and adds three controls with no ECC-2:2024 equivalent: brand protection, a licensed managed security operations subscription, and printer and imaging device security. The combined catalogue is 203 rows.

The control sets build on the 2018 controls and align with the international frameworks cited above, while sector rules such as the SAMA framework govern regulated industries in addition [44–46]. Strengthening SME cybersecurity is an explicit Vision 2030 objective, and the Personal Data Protection Law, the national data management regulations and the cloud-first policy determine much of what an entity must control [47–51]. Regional threat reports point the same way: the Kingdom sustained the Shamoon wiper campaigns of 2012 and 2016, whose operators moved through stolen credentials and reachable services rather than through novel exploitation. Both routes depend on internet-facing entry points that an external observer can identify [52, 53].

This capacity gap is measurable. Active commercial registrations exceeded 1.7 million in 2025 [54, 55], against an estimated 21,700 cybersecurity specialists nationally [56]. This is approximately one specialist per 78 registrations. The regulator's current assessment channels are declaration-based: a published self-assessment questionnaire, and the Haseen platform, which measures the compliance of registered entities and validates their email authentication configuration [57, 58]. Both channels work from what entities register and declare. None of these channels observes a decommissioned but still active subdomain, an orphaned endpoint, or an expired certificate. A dangling DNS record that resolves to a reclaimable resource is a documented subdomain-takeover vector, and no declaration-driven channel observes it [59]. The externally testable tier is not unique to these control sets: the United States already mandates externally observable configurations, HTTPS and DMARC among them, for federal agencies [60].

Research objective. The objective of this work is to design, implement and evaluate an automated toolkit that converts external observations into compliance evidence linked to control identifiers. We classify each control of a national cybersecurity catalogue by the evidence needed to assess it. The toolkit applies that classification: it produces evidence for each externally testable control from unauthenticated black-box observation, and it compares organisation-supplied declarations with that observed evidence. Four research questions follow.

- **RQ1.** Which controls of a national cybersecurity catalogue can unauthenticated external observation test, and what class of evidence do the remaining controls require?
- **RQ2.** Can the detector-to-control mapping on which the toolkit depends be derived automatically by keyword matching, dense retrieval or language-model inference, or does it require expert curation?
- **RQ3.** How accurately does the toolkit detect violations of the externally testable controls under a parameterised sensing model, relative to a predict-all baseline?
- **RQ4.** To what extent can the toolkit flag overstated declarations that observed findings contradict?

Contributions. This paper makes four contributions.

- **An observation boundary and a three-tier evidence model.** We define an explicit boundary over the facts unauthenticated black-box observation can access, and

classify all 203 rows of ECC-2:2024 and NCNICC-1:2025 by whether they are testable within that boundary and by required evidence class: externally testable, internal assessment, or declaration (Section 3.2). This answers RQ1 and provides the classification the toolkit applies.

- **An automated tiered-evidence toolkit.** We design and implement a four-stage toolkit that discovers internet-facing assets, includes a 31-detector suite, of which 15 are implemented as live inspectors, resolves each finding to the controls it supports through a curated mapping with automated integrity checks, and scores those controls (Section 3.3). Scoring follows the evidence-precedence rule: observed evidence takes precedence over declarations. A contradiction component checks organisation-supplied declarations against observed findings and flags those the findings contradict.
- **A synthetic evaluation cohort.** We release 100 synthetic Class B firms whose demographic parameters are drawn from national statistics, whose relative exposure patterns follow regional incident reports, and whose numerical exposure rates are explicit modelling assumptions (Section 3.4). A parameterised sensing model keeps ground truth separate from toolkit input and enables end-to-end evaluation where scanning real organisations without consent is not permissible.
- **An evaluation with baselines, sensitivity analysis and replication.** We show that generic text-similarity methods and twelve open-weight language models do not recover the curated mapping (RQ2). We then quantify violation-detection and contradiction-detection performance against explicit baselines and ceilings (RQ3, RQ4), test sensitivity to the modelling choices, and replicate on a structurally distinct synthetic cohort (Section 4).

The remainder of the paper is organised as follows. Section 2 reviews the related literature across five research strands and positions the present work against the closest systems. Section 3 presents the methods: the evidence-tier classification, the toolkit architecture and its components, and the construction of the synthetic evaluation cohort. Section 4 reports the results for RQ2 to RQ4, together with sensitivity and replication analyses, and traces a single firm end to end as a sample of toolkit output. Section 5 discusses the interpretation of the results and the limitations that constrain them, and Section 6 concludes.

2. Review of related literature

Five strands of research bear on the problem addressed here. This section reviews each in turn and then positions the present work against the closest systems; Table 1 summarises that comparison along the axes that separate them.

2.1. Compliance assessment driven by self-report

One group of systems relies entirely on organisational self-report. The risk-based compliance assessment system for ECC-1:2018 is the closest Saudi predecessor: a respondent answers a questionnaire, and four weighting schemes convert the responses into a compliance percentage, without independent observation or supporting evidence [14]. A 2026 tool for Jordan's national framework repeats this design, extended with additional mappings and a richer scoring model, and remains driven entirely by self-reported status [16]. Maturity assessments for Saudi higher education follow the same pattern [15], as do compliance assessment systems published in this journal [17] and data-driven assessment tools developed for the European SME population [18]. A more recent framework removes the questionnaire entirely: it recommends which controls an organisation should adopt from a set of standards, using a self-declared maturity level as its only organisational input. The tool uses no other information about the organisation [19]. A systematic review of the SME compliance literature likewise finds the same reliance on self-reported status across the field [20]. None of these tools compares the self-report with independent observation.

2.2. Cooperative validation of self-reports

A second group checks self-reports using evidence from within the organisation. Chidukwani and colleagues assessed eight SMEs through expert site visits, interviews, artefact review and an internal vulnerability scan executed from a laptop on each firm's own network [13]. They benchmarked each firm against the NIST Cybersecurity Framework and found that self-reporting alone overstates posture. The same group earlier identified the barriers that prevent small firms from acting on security guidance [9] and surveyed their available security measures [10]. Studies of Saudi small firms report the same shortage of staff and process [61, 62], and research documents the cost of this gap [12]. A comparative study of Saudi and United Kingdom SMEs also finds low awareness of the national control sets [63]. Even carefully designed surveys face the same limitation at scale, since a well-conducted survey records what respondents report rather than what their organisations expose [11].

2.3. External observation without control-level mapping

A third body of related work observes organisations from outside but stops short of control-level classification. Rahaman and colleagues used external scans to test payment card certification and found that some sites were certified even though external scans showed non-compliance, which motivates a tiered treatment of what external observation can and cannot test [40]. Externally observable mismanagement predicts breaches without an organisation's cooperation [32, 64], and commercial security-rating services convert such observation into scores, although independent comparison shows that scores differ across providers and are difficult to audit [39]. The cyber-risk literature treats external observation as a useful but limited predictor [36], links observed posture to breach likelihood [35], and maps it to the controls cyber insurers ask applicants about [65]. National assurance schemes go further and pair a questionnaire with an external scan, as Cyber Essentials Plus does, but issue an organisation-wide pass or fail result

Table 1: Comparison of the closest systems by evidence source, framework, automation, population, reproducibility and output. Among the compared systems, this is the only one that produces per-control evidence from external observation of a national framework and releases the artefacts required to reproduce the results.

System	Evidence	Framework	Auto.	Population	Repro.	Output
RC2AS [14]	self-report	ECC-1:2018	no	1 firm	no	score
Jordan JCCT [16]	self-report	Jordan JNCSF	partial	1 firm	no	score
Saudi HEI [15]	self-report	ECC-1:2018 aligned	no	survey	no	maturity level
VCPAF [13]	scan + interview	NIST CSF	no	8 firms	no	posture report
PCI certification [40]	external scan	PCI DSS subset	yes	1,203 sites	yes	pass or fail
Security ratings [39]	external scan	none	yes	many firms	no	opaque score
Cyber Essentials Plus [41]	form + scan	UK baseline	partial	certified firms	no	pass or fail
AI framework [19]	self-declared	multi-standard	yes	3 scenarios	no	control list
This work	external observation	ECC-2:2024 + NCNICC-1:2025	yes	100 synthetic firms	yes	per control

rather than producing per-control evidence across a framework [41]. External observation is an established signal, but we have identified no prior system that maps it to the individual controls of a national standard.

2.4. Internet measurement at scale

Internet measurement provides the methods the toolkit uses. Fast scanners enumerate the reachable address space while search engines index the services that respond [24, 25]. These methods have measured email authentication adoption across the web [26, 27], predicted organisational risk from externally visible features [33, 34], and matured into attack-surface management, which focuses on the unmanaged internet-facing footprint [28–31]. Measurement surveys describe the capabilities and limitations of these methods [37, 38], among them the sampling bias of public top lists [66]. These methods are now being applied at national scale in the region: web security has been measured across Türkiye at national scale in this journal [67], and the privacy practices of 723 popular Saudi websites have been scanned bilingually [68]. To our knowledge, no national-scale measurement of Saudi security posture yet exists.

2.5. Machine-readable control catalogues and continuous compliance

The standards have also been represented as structured data, and this structured representation underpins the classification. NIST publishes control catalogues as structured documents through OSCAL and its predecessor SCAP, and maintains cross-framework mappings as reference data [4, 21, 22]. Knowledge-based methods use these representations to check policy against configuration [23], continuous compliance systems evaluate cloud controls against live evidence on a schedule [5, 6], and framework evaluations compare standards for regulatory fit [7, 8]. These systems encode control text as data and reason over it, and their evidence comes from within the organisation's administrative boundary, whether from internal configuration or from cloud-native telemetry.

2.6. Positioning

Together, these five strands reveal a specific gap: the mapping from external observation to the controls of a national framework. We address it by defining an explicit observation boundary over the complete ECC-2:2024 and NCNICC-1:2025 control sets and classifying every control by that boundary and by evidence tier. An automated toolkit implements the classification, producing per-control evidence and flagging contradicted declarations, and we evaluate it on a synthetic population of 100 firms rather than a single case. We release all artefacts required to reproduce the results. The distinction between observation and elicitation is established [32, 69]. The new elements are the control-level classification of ECC-2:2024 and NCNICC-1:2025, the evidence tiers beneath it, and the toolkit that implements the classification.

3. Methods

3.1. Design objectives and overview

The toolkit is designed around four objectives that follow from the requirements identified in Section 1. It must operate without credentials, agents or organisational cooperation, so that organisations without security staff can use it. It must produce evidence linked to control identifiers rather than to exposures, so that its output is usable in formal assessment. It must state explicitly which controls it can test and which it cannot, rather than returning a single composite score. And it must incorporate organisation-supplied evidence for the controls it cannot test, so that most controls, which lie beyond the scan, are not left unaddressed.

Three components implement these objectives, and the remainder of this section describes each in turn. Section 3.2 presents the evidence-tier classification, which partitions the control sets and provides the mapping the toolkit uses. Section 3.3 specifies the toolkit itself: its processing stages, its detector suite, its mapping component, its scoring rules, its contradiction component and the adversary model under which it operates. Section 3.4 describes the synthetic evaluation cohort and the sensing model applied to it. Together they enable end-to-end evaluation without scanning real organisations.

3.2. The evidence-tier classification

3.2.1. The observation boundary

The classification uses a defined observation boundary and a rule that classifies each control against it. The boundary is the set

of facts an unauthenticated external observer of an organisation's internet-facing assets can determine. Observable facts include published DNS records, TLS handshake parameters, responsive ports, web response headers, and disclosed service names and versions. A control is externally testable when these observations determine, for each asset the scan discovers, whether the asset satisfies the control. At organisation level, the evidence is one-directional: a finding demonstrates a violation, whereas the absence of a finding does not prove compliance, because the scan may not discover every asset. Where observation is insufficient, the organisation must provide its own evidence for the control.

3.2.2. Evidence tiers

The controls the scan cannot test require different types of evidence, so the classification uses three tiers. Some cannot be tested externally but can be supported by a structured internal record: a device management console, a backup restoration test, a penetration test report, or a log retention configuration. Other controls concern a governance committee, a personnel screening policy, or physical access to a server room. They rest solely on a declaration by an accountable individual, because neither external observation nor a routine internal record determines whether they operate. The distinction is whether the control's routine operation produces a directly inspectable record that allows a reviewer to determine whether the control operates. Routine operation of a Tier 2 control produces such a record: an export, a log or a test report a reviewer can inspect. It is Tier 3 when any document the organisation could produce is an account written by the organisation about its own practice, so assessment ultimately depends on attestation. Minutes, for example, may show that a committee met, but they do not by themselves establish that the governance process operates. An organisation planning its compliance effort and cost needs to know which controls it can support through internal assessment and which require a signed statement. Tier 1 is externally testable, Tier 2 requires internal assessment, and Tier 3 requires a declaration. Tiers 2 and 3 together form the attestation-required remainder, the controls for which the organisation must provide evidence. The primary distinction in this paper is between Tier 1 and that remainder.

3.2.3. Distribution across the catalogue

Externally testable rows are a minority of the catalogue, and their distribution across the standard is uneven. Table 2 reports the split: 23 of the 203 rows are Tier 1, 72 are Tier 2, and 108 are Tier 3; Table 3 gives illustrative rows from each tier. Figure 1 shows the imbalance. All 23 externally testable rows fall within the defence domain. The governance, resilience and third party domains, together with the NCNICC-1:2025 additions, contain none at all. Outside the defence domain, external observation provides no evidence and the organisation must supply its own. The size of the externally testable tier is itself a result, because it defines the limits of external observation across the catalogue.

Table 2: Classification of the 203 ECC-2:2024 and NCNICC-1:2025 control rows by the type of evidence required. The externally testable tier is 11.3% of the catalogue.

Tier	Controls	Share (%)
Tier 1: externally testable	23	11.3
Tier 2: internal assessment	72	35.5
Tier 3: declaration	108	53.2
Total	203	100.0

Table 3: Illustrative rows of the classification across the three tiers, keyed by control identifier. Control text is NCA copyright and is not reproduced.

Control	Tier	Detectors	Class B
2-8-3-3	externally testable	6	mandatory
2-4-3-5	externally testable	4	mandatory
2-15-3-3	externally testable	10	recommended
2-9-3	internal assessment	0	mandatory
1-9-3-1	declaration	0	not in scope
2-14-3	declaration	0	mandatory

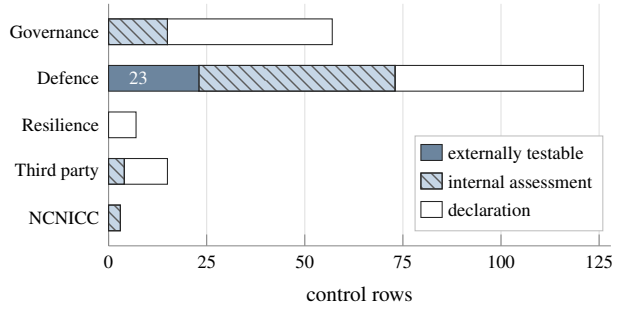


Figure 1: Composition of each domain by evidence tier. All 23 externally testable rows fall within the defence domain; the other domains contain none.

3.2.4. Granularity of the count

The 23 Tier 1 rows consist of 16 subcontrols and the 7 parent controls that contain them. We report the count at both levels of granularity, because a main control containing a single externally testable subcontrol is itself only partly testable from outside. At the subcontrol (leaf) level, 16 of the 92 subcontrols are externally testable; that is the stricter figure.

3.2.5. Detector-to-control mapping

The detector specification contains 31 detectors, mapped to exactly the 16 Tier 1 subcontrols. Figure 2 maps the detectors to the controls.

Ten of the 31 detectors map to a single control, 2-15-3-3, which specifies the use of secure protocols for external web applications; the web header and web application surface detectors all fall under it. That concentration is the most contestable interpretation in the classification, and 2-15-3-3 is the control most likely to attract expert disagreement. We therefore retain this contestable mapping rather than reassign detectors to increase apparent coverage.

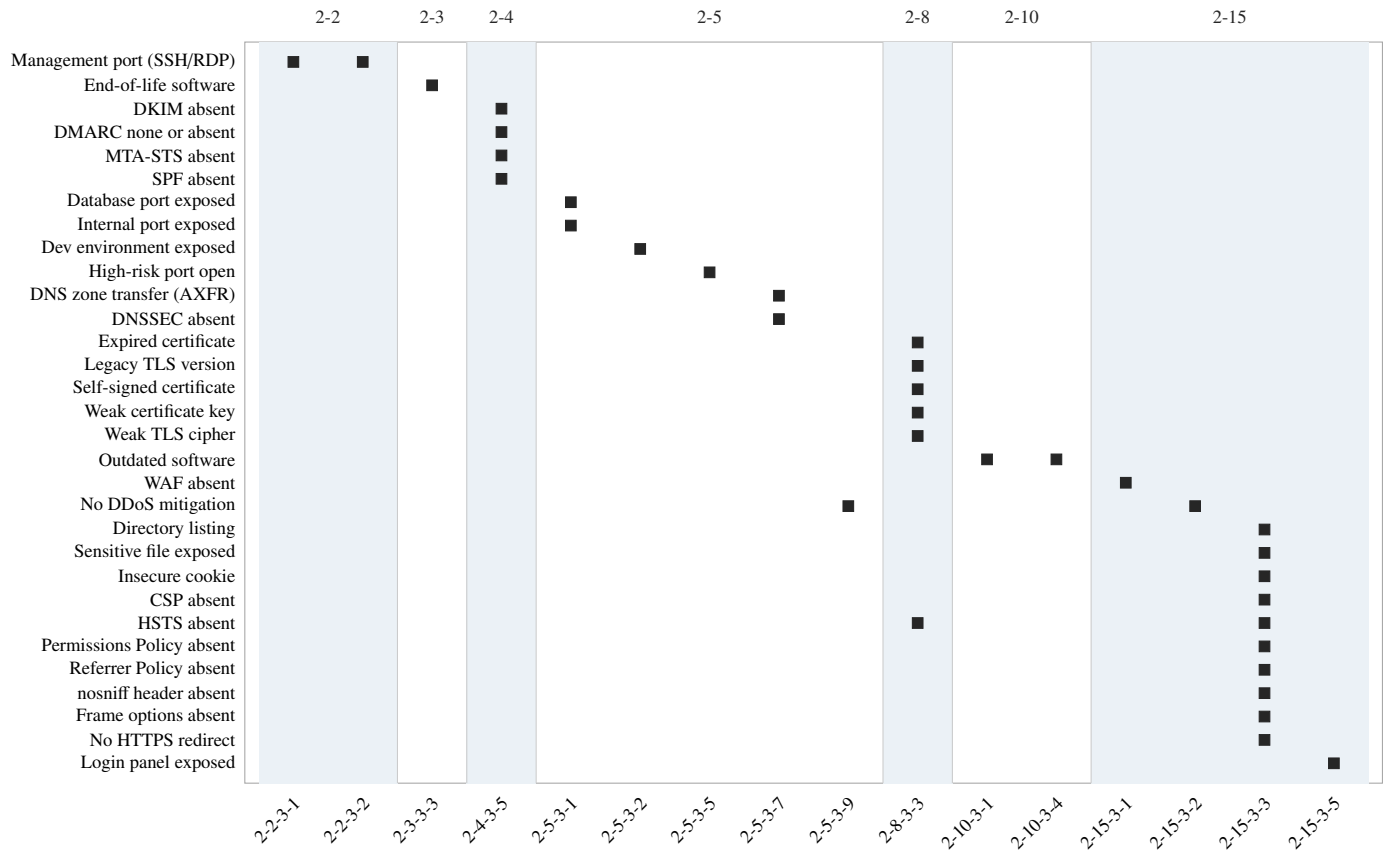


Figure 2: Each detector maps to the subcontrols it supports. The 31 detectors map to exactly the 16 externally testable subcontrols, and ten resolve to control 2-15-3-3, the concentration discussed in Section 5.

3.2.6. Curation procedure

Controls fall outside the externally testable tier when they leave no trace an external observer can read. Governance controls, personnel security, physical protection, backup and recovery, incident response, and internal patch cycles all do so, and must therefore be classified as internal assessment or declaration. Each tier was assigned from the standard's text. For 42 of the 200 ECC-2:2024 rows, an initial pass over the control text assigned a provisional default tier. We reviewed all 42 against the control text and the NCA implementation guide; review confirmed the default in every case, no tier was moved, and each decision is recorded in `docs/DECISIONS.md`. The remaining rows were assigned directly from explicit wording in the control text, and the three NCNICC-1:2025 additions fall in Tier 2 using the same criteria. Section 5 states the status of this curation.

The next section specifies the toolkit that applies this classification: it maps scan findings to controls and scores the controls to which those findings map.

3.3. Toolkit architecture

3.3.1. Processing stages

The toolkit converts an external scan into per-control evidence in four stages, and keeps the stages separate to allow independent audit. It discovers an organisation's internet-facing assets, inspects each asset with the detectors, maps every finding to the controls it supports through the classification, and scores the mapped controls. Figure 3 illustrates the stages using one worked firm, eval-000, which Section 4.7 examines in full.

3.3.2. The detector suite

The 31 detectors read only data available within the observation boundary. They cover response headers, transport security and certificates, open and management ports, email authentication records, DNS security, and exposed development or administrative surfaces. The detector definitions follow established external-measurement practice, and open-source tools already probe the same surface for a single organisation [70–72]. The transport and certificate detectors build on large-scale TLS and HTTPS measurement [73], and the email authentication detectors read the SPF and DMARC records those standards define, while DKIM is probed through its published selectors and therefore remains selector-limited [74–76]. Fifteen of the 31 detectors are implemented as live inspectors in the released toolkit. The remaining 16 are specified in the detector catalogue with their control mapping but are not yet implemented; each detector also has a CVSS v3.1 base severity [77]. Sections 4 and 5 report how much of the evaluation depends on these specified but unimplemented detectors.

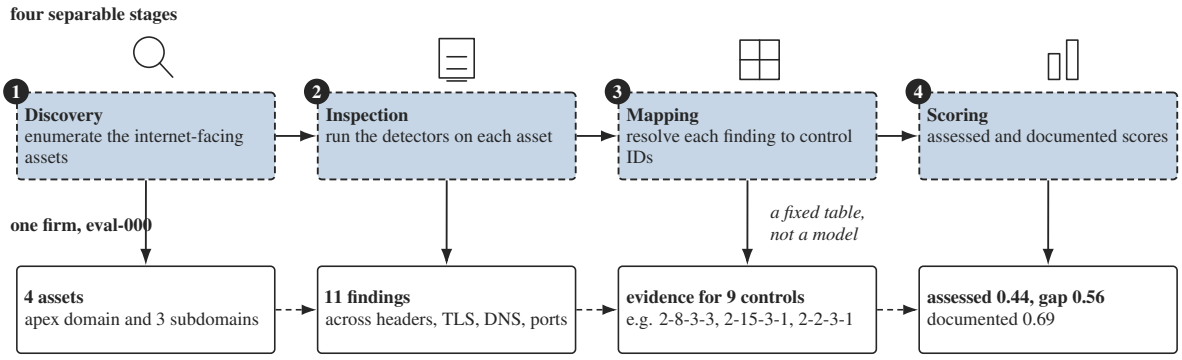


Figure 3: The toolkit converts an external scan into per-control evidence in four separable stages, traced here on the worked firm eval-000. Separating the stages keeps each stage auditable.

3.3.3. The mapping component

Mapping is performed by a fixed table rather than a learned model. Each detector stores the identifiers of the controls it supports, checked against the control set at toolkit start-up, so the toolkit cannot output an identifier the standard does not define.

3.3.4. Coverage scoring

The toolkit reports two scores, because they answer different questions. When evidence for a control is available at more than one tier, the evidence-precedence rule applies: the highest-priority evidence determines the control's status, in the order Tier 1, then Tier 2, then Tier 3. A Tier 1 finding therefore takes precedence over a Tier 3 declaration. The assessed score is the share of the 16 externally testable subcontrols on which the scan observes no violation within the discovered estate. It summarises the observed external posture. The externally attributable gap equals one minus the assessed score. This score is limited by the one-direction rule: a subcontrol without an observed violation has not been shown compliant, because the scan may not discover every asset.

The documented score is the share of evidenced controls whose highest-priority evidence records no gap. Its denominator covers more controls: the externally testable subcontrols plus every control the organisation supports with Tier 2 and Tier 3 records. Only evidence for off-surface controls expands this denominator, since a declaration on an externally testable control is determined by that control's Tier 1 evidence and adds nothing. The documented score summarises the organisation's documented position, and it relies on self-report wherever no finding provides an external check. A control outside the externally testable set for which the organisation supplies no evidence is excluded from both scores; it remains an unresolved obligation, which the toolkit counts separately (Section 4.7). Reported results use uniform weights, so that every control contributes equally. An expert weighting is supported; the toolkit accepts elicited weights only when they pass a consistency check, and this paper does not use them.

3.3.5. Contradiction flagging

The evidence-precedence rule lets external observations check declarations. When an entity declares a control compliant at Tier 3 and a Tier 1 finding conflicts with that declaration, the toolkit raises a contradiction flag. The scan cannot test the 180 attestation-required rows. It can, however, flag a potentially overstated declaration about a control the scan can observe. Because findings can be false positives, each flag indicates a potential contradiction and requires review; Section 5.3 quantifies the rate. Figure 4 shows the precedence rule and the contradiction branch.

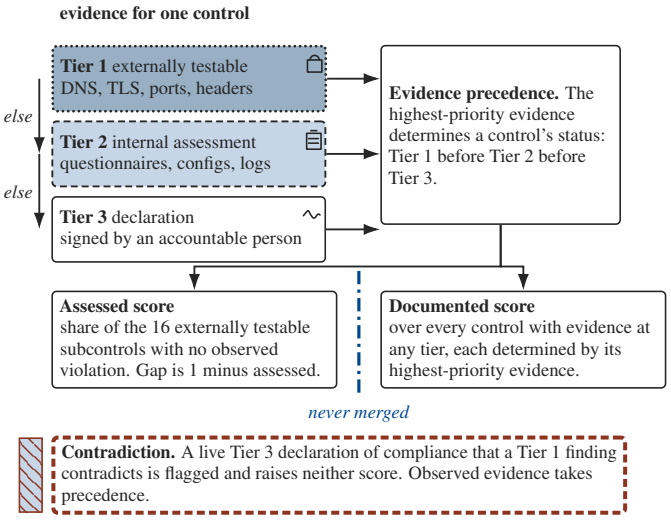


Figure 4: Evidence is scored per control under the evidence-precedence rule: the highest-priority evidence determines each control's status, so observed evidence takes precedence over a declaration and the scan becomes a check on what an organisation claims.

3.3.6. Adversary model

The adversary is an organisation that overstates its posture. It controls its declarations and its externally visible configuration, and it may know that a scan will occur. Its objective is to appear compliant. It does not control the scanner or the third-party records the scan reads. The observation boundary defines what the organisation cannot conceal and what it can. Everything inside the observation boundary is read directly by the observer rather than taken from the organisation's claims, although a self-reported version banner can still be false (the second evasion

below). Governance, personnel, physical and backup controls leave no external trace (Section 3.2) and lie beyond the scan's reach.

Three evasion strategies follow directly from the boundary.

1. An organisation that declares nothing leaves no declaration for a finding to contradict: the check can evaluate only declarations that have been submitted. Because the declaration is itself mandatory, its absence is a failure the regulator can see.
2. Banner spoofing and finding suppression are effective evasions, and the version and software detectors are vulnerable to both.
3. Blocking probes affects the whole score rather than a single finding: it prevents the toolkit from observing violations, so an organisation that blocks the scanner raises its own assessed score. The assessed score should therefore be treated as an upper bound and read alongside scan coverage. A deployment should report the unreachable share of the estate alongside the score. That share is the field counterpart of the reachable, throttled, content-delivery-network (CDN) fronted and blocked shares modelled in Section 3.4.

The evaluation models a non-adaptive declarer, so the contradiction result measures detection of non-adaptive overstatement rather than of an organisation adapting to the scan. The check further assumes a trusted scanner used with the organisation's consent. The check fails if an organisation scans its own estate and edits the output, so the operator and the integrity of the scan matter as much as the observation itself. The trusted components are few: the scanner, the operator who runs it, and the third-party records used by the scan, namely append-only certificate transparency logs and registration data held by domain registries. Tier 2 evidence is organisation-supplied and is not externally verified, so the documented score inherits the organisation's self-report except where a Tier 1 finding contradicts it.

3.4. The synthetic evaluation cohort

3.4.1. Demographic parameters

We evaluate the toolkit on 100 synthetic Class B firms. The generator's parameters fall into three categories. Firm size, sector mix, estate size and domain mix are drawn from Saudi national statistics. The relative exposure patterns of four finding categories follow regional incident reports. Every numerical exposure rate is a declared modelling assumption, as is the posture distribution, which no Saudi source measures by sector. We use synthetic firms because scanning real Saudi organisations without consent is out of scope.

Within Class B, 89% of firms are small and 11% are medium. This split is taken from the Monsha'at register for the fourth quarter of 2023 [78], with firms sized by the official Monsha'at employee bands [79]. The sector mix follows the Monsha'at stock distribution and the General Authority for Statistics (GASTAT) establishment survey, so that retail and construction dominate as they do in the economy [80, 81]. Estate size depends on the firm's size band and is anchored to GASTAT figures on internet and cloud adoption [82], and the domain mix reflects the low share of Saudi firms holding a national domain [83]. A small fraction of those national registrations carry Arabic-script labels, and the generator represents such domains in Punycode, the encoded form an external scanner reads on the wire. Appendix A lists the population priors, the incident-informed exposure weights, and the detection rates with their sources. The GASTAT ICT access survey indicates the share of the population an external scan can reach: 32.1% of Saudi establishments operate a website while 95.0% use email. The cohort therefore represents the web-exposed third of the population, and the email authentication detectors apply to nearly all of it [82].

Regional incident reports inform the exposure model. No Saudi source measures SME exposure by category, so all numerical exposure rates remain declared assumptions, with upper bounds informed by the population-level anchors of Appendix A. Regional reports inform only the relative direction and ranking of four categories. Saudi incident-response and security-operations reporting indicates that most initial access occurs through internet-facing systems: exploited public-facing applications and valid accounts rank far ahead of phishing, and brute force is the most detected attack type [84]. That source is vendor telemetry and reports rank order rather than sample sizes. Global data show a similar pattern: exploitation of VPN and edge devices rose eight-fold as an initial-access vector in a single year [85]. The vulnerabilities observed to be exploited in the Kingdom affect internet-facing enterprise applications and remain exploited years after disclosure; a SharePoint vulnerability disclosed in 2019 (CVE-2019-0604) still ranked among the most exploited of 2023 [84]. Regional ransomware victims are concentrated in mid-market manufacturing and construction firms [86].

The generator therefore assigns higher probabilities to those four categories: management and high-risk ports, outdated software, end-of-life software, and reachable authentication surfaces (Table A.10). The largest increases apply to construction and manufacturing, the sectors identified in regional reports. Transport receives the same increase as a declared modelling choice. The remaining categories are left unchanged, for the following reasons. Regional reports provide no data on headers, transport security and development environments, and they provide only qualitative evidence about web application firewalls (WAF) and denial-of-service (DDoS) protection. On DNS security the national zone is signed yet scarcely used, which the high prior already reflects. And on email authentication the available evidence does not support assuming below-global adoption, since large Saudi organisations already publish enforcing records.

These sources are used only to set the generator's priors.

3.4.2. Generative procedure

Each firm is a deterministic function of its index and one global seed, so that the same seed regenerates the cohort and every value derived from it. The generator draws in a fixed order. A size band, small or medium, determines the estate size and

Algorithm 1 Generation of one synthetic Class B firm. Detect observes each true issue with the product of the detector's sensitivity, the firm's scanner competence factor, and the host's acquisition factor, and adds the false positives of Section 3.4.3.

Require: firm index i, global seed s

$r \leftarrow \textsc{Random}(s + i)$ ▷ one seed per firm makes the cohort deterministic
$band \leftarrow \textsc{Draw}(r, \{\text{small} : 0.89, \text{medium} : 0.11\})$
$sector \leftarrow \textsc{Draw}(r, \text{eight sector prevalences})$
$posture \leftarrow \textsc{Draw}(r, band.\text{posture shares})$
$apex \leftarrow$ domain under a weighted TLD; $hosts \leftarrow \{apex\} \cup$ subdomains from $sector$'s pool, count set by $band$
$true \leftarrow \emptyset$
for all host $h \in hosts$ and finding category c **do**
  if $\textsc{Bernoulli}(r,\ sector.w[c] \times posture.\text{mult})$ **then**
    for all detector k in category c **do**
      add $k@h$ to $true$ with probability p_k
    end for
  end if
end for
$observed \leftarrow \textsc{Detect}(true)$ ▷ $\text{sens}_k \times \text{comp} \times \text{acq}_h$
$controls \leftarrow \textsc{Map}(observed)$ ▷ the fixed classification
$evidence \leftarrow$ Tier 3 claims on the scanned controls, some overstated, plus Tier 2 and Tier 3 evidence off surface
return $(hosts, true, observed, controls, evidence)$

the posture priors. A sector is drawn by prevalence. A posture level of strong, moderate or weak is then drawn from the band's shares. An estate is then constructed as an apex domain plus a number of subdomains determined by the size band, sampled from a sector-specific label pool. A retail firm therefore exposes a storefront, a checkout and the e-invoicing integration every VAT-registered entity operates; a construction firm exposes a project portal. The generator then injects the firm's true external issues. For each host and each finding category it samples presence using a Bernoulli trial, with probability the sector weight for that category scaled by the posture multiplier. If the category is present, each detector in the category is included with a detector-specific probability. The result is the firm's unobserved true issue set. This is the ground truth, which the toolkit never receives. Finally, the generator draws the firm's declarations and off-surface evidence. It draws Tier 3 claims on the externally testable controls, overstating a declared share of them. It then draws the Tier 2 and Tier 3 records the firm holds for controls no scan can assess, with coverage that increases as posture improves. Algorithm 1 states the procedure in full.

3.4.3. The sensing model

A modelled sensing layer converts the unobserved issues into the findings supplied to the toolkit, because an external scan does not observe every issue. The observation probability for each true issue is the product of three factors. The first is the detector's sensitivity, set per detector between 0.70 and 0.99. The second is the firm's scanner competence factor, drawn per firm between 0.97 and 1.00, which models the run-to-run acquisition quality of a single authorised scan of that firm. The third is a per-host acquisition factor determined by the host's scan condition. The DNS-resolved apex detectors are the exception: published DNS records remain readable regardless of the host's scan condition, so neither the competence nor the acquisition factor applies to them. A host is reachable, throttled, behind a CDN, or blocked, with shares 0.85, 0.09, 0.05 and 0.01 respectively. A CDN suppresses origin-facing detectors at a factor of 0.40 but barely affects edge-facing ones at 0.95, and a blocked host has a low acquisition factor of 0.20.

Scan condition and competence are drawn independently of posture, so any correlation between a firm's exposure conditions and the quality of its administration is excluded from the model. A CDN fronting a better-managed estate is one such correlation. The stress test that comes closest to bounding this effect sets every low-confidence detection rate to its pessimistic extreme at once; these are the ambiguous WAF, DDoS, DKIM, banner and login signals. Under it, the violation-detection task of Section 4.3 holds at F_1 0.872.

The sensing layer generates false positives on 16 of the 31 detectors, at rates between 0.02 and 0.24 per host. The other 15 detectors read published values deterministically: DNS records, response headers, certificate fields and email authentication records. By construction they carry no false-positive rate. Under the inspector contract, a live inspector reports absence only on a definitive read. A timed-out or failed lookup produces no finding rather than an absent record, leaving obstruction and unreachability to the acquisition factor above. These 15 zero-false-positive detectors are not the same 15 that have a live inspector; the two groups overlap but are not identical.

3.4.4. Leakage control

The true issue set and the observed finding set are kept in separate data structures, and an automated test checks the serialised toolkit input for ground-truth fields, preventing label leakage into what the toolkit scores. The deterministic reads and the high-reliability probes are calibrated against controlled targets and one authorised scan of author-owned infrastructure; published figures instead provide bounds for the nine analyst-judgement rates (Section 5.1 gives the split). Appendix A reports every rate with its source or modelling status.

4. Results

4.1. Experimental setup

Three experiments address RQ2 to RQ4 in turn. All three are conducted on the 100-firm synthetic cohort at a fixed seed. A robustness analysis tests the conclusions against the modelling choices, and every reported value can be regenerated from the released artefacts. The mapping, the toolkit, the cohort and every evaluation output accompany the paper and will be made publicly available upon acceptance. Throughout, the positive class is a control the firm truly violates; Section 4.3 defines the measures. Table 4 collects the principal results. Table 5 reports observed findings by category; web-header findings dominate because ten detectors map to the single control 2-15-3-3.

Table 4: Principal results on the 100-firm cohort. The implemented-only rows restrict the violation-detection task of Section 4.3 to the 7 of 16 controls covered by an implemented inspector and are paired with a matching predict-all baseline. Every value is transcribed from the released artefacts and gated by `verify_numbers.py`.

Quantity	Value
Violation detection (RQ3)	
Violation-detection micro precision	0.892
Violation-detection micro recall	0.925
Violation-detection micro F_1	0.909
Violation-detection macro F_1	0.884
Violation-detection Matthews correlation	0.779
Violation-detection balanced accuracy	0.887
Violation-detection F_1, implemented only	0.943
Predict-all baseline, implemented only	0.758
Predict-all baseline F_1	0.732
Perfect-mapper expectation F_1	0.910
Mapping integrity F_1	1.000
Contradiction detection (RQ4)	
Contradiction precision	0.853
Contradiction recall	0.947
Contradiction F_1	0.898
Mapping baselines (RQ2)	
Best generic mapper F_1	0.576
Single-pick ceiling F_1	0.916

Table 5: Observed findings by category over the cohort (1,857 findings). Web-header findings dominate, as discussed in Section 5.

Category	Findings
web headers	647
ports	266
TLS	224
WAF	185
software	124
email	112
end of life	109
auth	77
DNS	61
DDoS	41
dev environment	11

4.2. Automatic derivation of the mapping (RQ2)

We first test whether generic text-similarity methods reproduce the curated mapping. They do not. This tests portability: extending the classification to another control set requires expert time only if no generic method can derive the mapping. We compare keyword matching, lexical and semantic embedding, and twelve open-weight language models with the curated mapping, weighting each detector by its finding frequency in the cohort. The models range from 7 to 32 billion parameters and include both general and Arabic-specialised systems. All are evaluated zero-shot (Appendix B). The best of them, a 12-billion-parameter model, reaches F_1 0.58 (Figure 5).

The 0.916 line in Figure 5 is a ceiling on single-pick scoring rather than on the mapping itself: some detectors have two reference control labels, so naming one control per detector cannot reach 1.000. Lifting that constraint does not materially improve the generic mappers. A lexical baseline allowed to return several controls per detector still scores only F_1 0.19 against the curated 1.000, and no language model under the same multi-label setting exceeds 0.32 (Appendix B). This experiment measures agreement with the curated reading rather than its correctness: a superior mapping would score lower here. We test zero-shot prompting because a few-shot prompt containing mapping examples would reveal the target schema, and fine-tuning would require the labelled mapping being evaluated. A constant mapper that assigns the single most frequently supported control to every detector scores F_1 0.319 and forms the constant-mapper baseline drawn as the dotted line in Figure 5; that control accounts for a finding-weighted modal share of 0.348.

4.3. Violation detection (RQ3)

The violation-detection task compares the toolkit's per-control predictions with the generator's ground truth. The positive class is a control the firm truly violates. A control with neither a predicted nor a true violation is a true negative, which micro-F_1 ignores but the Matthews correlation and balanced accuracy reported below both count. On the cohort the toolkit achieves a modelled micro-F_1 of 0.909 (95% bootstrap interval 0.891 to 0.925), with precision 0.892 and recall 0.925; the macro-F_1 is 0.884. This result is an integrity check on the toolkit rather than a measurement of field scanner performance: mapping integrity is 1.000 by construction, so the violation-detection result primarily reflects the modelled detection rates. Restricted to the controls covered by an implemented inspector, violation detection is F_1 0.943 over 7 controls, against a matching predict-all baseline of 0.758.

The predict-all baseline labels every control as violated. It scores 0.732 at the cohort's base rate of 0.577, so the toolkit exceeds it by 0.177. That margin is not an artefact of the base rate: over the 1,600 firm-by-control decisions the toolkit attains a Matthews correlation of 0.779 and a balanced accuracy of 0.887. Both measures use the full decision table, on which the predict-all baseline equals 0 (the convention for a constant predictor) and 0.5 respectively, regardless of the base rate.

A perfect mapper facing the same modelled rates reaches F_1 0.910 on the expected detection counts. Because F_1 is not linear in those counts, individual runs vary around that expectation rather than being capped by it. Across thirty seeds the mean F_1 is 0.903 with a standard deviation of 0.009, and four of the thirty seeds exceed the expectation. The reported seed at 0.909 lies between the multi-seed mean and that expectation.

Most uncertainty arises from the modelled detection rates. A 2,000-draw Monte Carlo ensemble over them gives a 95% predictive interval of 0.896 to 0.927, comparable to the 95% bootstrap interval of 0.891 to 0.925 obtained by resampling the 100 firms. The detectors whose false-positive rates contribute most to this uncertainty are DDoS protection, DKIM and the exposed login panel, so calibration effort should begin there.

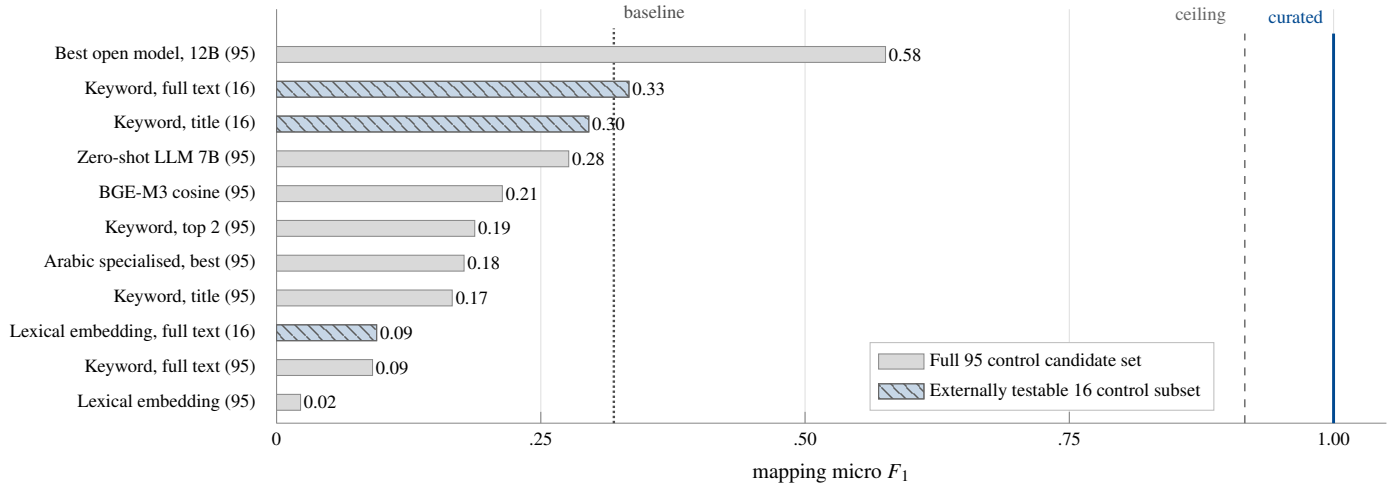


Figure 5: Generic mappers scored against the curated mapping over 95 candidates (the 92 ECC-2:2024 subcontrols plus the 3 NCNICC-1:2025 additions). Hatched bars restrict the candidates to the 16 externally testable subcontrols; the dotted line is the constant-mapper baseline. The best open-weight model reaches F_1 0.58 against the 0.916 single-pick ceiling.

Figure 6(a) gives the per-control precision and recall, and marks the controls covered only by specified but unimplemented detectors.

Precision varies with posture, rising from 0.66 on strong-posture firms to 0.94 on weak-posture firms, because a firm with few true findings yields a higher ratio of false positives to true positives. The aggregate is therefore weighted towards the 54 weak-posture firms. The 13 strong-posture firms present the operationally harder case of detecting a well-secured firm's residual gaps. In that subgroup violation detection falls to F_1 0.76, against a predict-all baseline that falls further, to 0.459. The posture effect is large: Cliff's δ, a nonparametric effect size comparing two distributions, is 0.82 for the per-firm F_1 of weak-posture versus strong-posture firms.

4.4. Contradiction detection (RQ4)

Overstated declarations are flagged at a contradiction F_1 of 0.898. The cohort contains 1,871 declarations, of which 1,597 are active and 274 lapsed or not yet effective. Half of the active declarations concern controls the scan cannot observe: 806 of the 1,597 fall outside Tier 1, where no finding can confirm or contradict them. That leaves 791 declarations that an observed finding could contradict. The toolkit nonetheless receives all 1,871 and correctly raises no flags for the lapsed declarations or for those outside Tier 1. An overstatement is a case in which a firm declares compliance with a control it truly violates. Among the 791, the toolkit flags overstatements at F_1 0.898 (95% bootstrap interval 0.871 to 0.922), with precision 0.853 and recall 0.947. Missed findings limit recall and false positives limit precision, the same two rates that limit violation detection.

This is the only measure reported here that tests what an organisation declares, and the counts above show its limited scope: only declarations concerning externally testable controls can be checked at all. A finding can contradict a claim, but the absence of a finding cannot confirm one, since the scan may not observe every asset. The evaluated declarer is non-adaptive; Section 5 discusses this limitation.

4.5. Sensitivity analysis

The modelling choices do not change the main conclusion. Shifting the population priors by 20% keeps the violation-detection F_1 within 0.881 to 0.930. A uniform shift of every detection rate by 0.10 constitutes a worst-case stress, since it also shifts the deterministic reads, whose rates would not in fact move. Under it the violation-detection F_1 spans 0.844 to 0.966, driven by the analyst-judgement rates. Varying the declared overstatement rate from 0.30 to 0.70 changes the contradiction F_1 between 0.840 and 0.903, so the contradiction result is reported at the declared rate of 0.50 rather than at the value producing the best result. Figure 6(b) reports the full sweeps. Table 6 lists the evaluation pitfalls and the corresponding safeguards, following the checklist for machine learning in security [87]. Among these is the separation of the ground truth from the toolkit input, which an automated leakage test enforces.

4.6. Replication on a structurally distinct cohort

Because the violation-detection figure restates the modelled detection rates, it may depend on the structural choices of a single generator. We therefore constructed a second cohort, cohort B, of 100 firms using a deliberately different specification. It uses indexed hostnames with positional roles, an unobserved per-host neglect variable in place of the category-level correlated draws of Section 3.4.2, a continuous posture variable in place of the three bands, and four independent sectors in place of the sector-by-category weights. We did not tune cohort B to match cohort A, and the 31 detectors, the mapping, the detection rates and the contradiction mechanism were held fixed. Its exposure structure also predates and does not include the

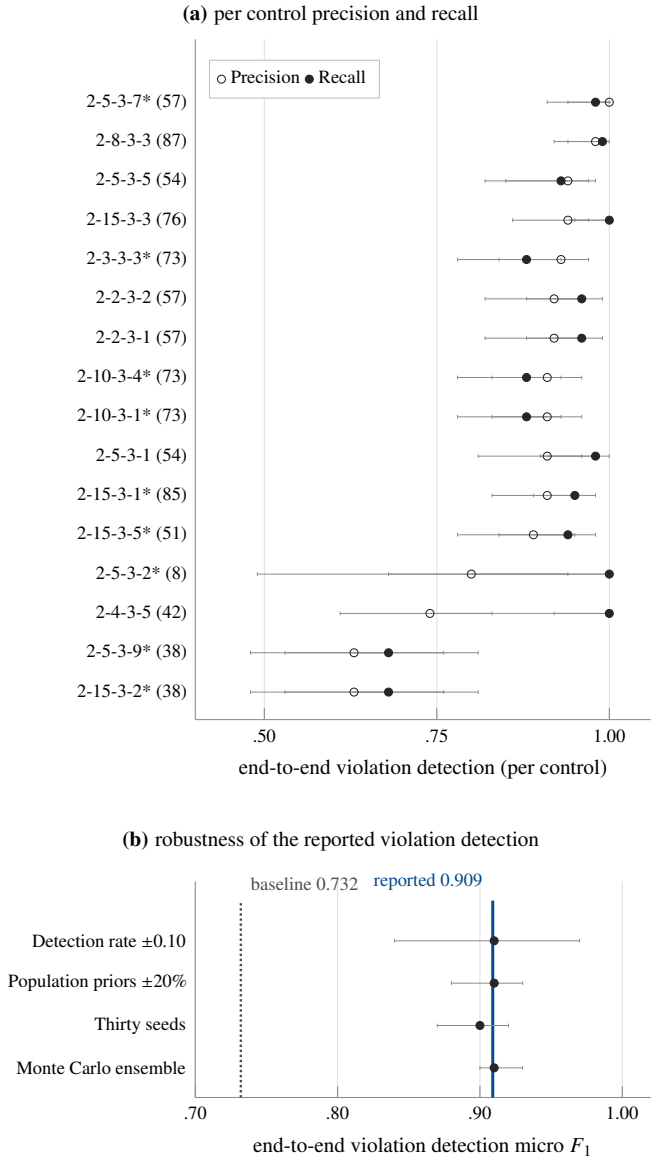


Figure 6: Panel (a) gives per-control precision and recall over the 100-firm cohort; controls covered only by specified but unimplemented detectors are marked with an asterisk, since they contribute to the violation-detection result. Panel (b) shows the violation-detection F_1 under four modelling sweeps, with every band clearing the predict-all baseline of 0.732 and close to the reported 0.909.

incident-informed reweighting of Section 3.4.1, so the transfer does not depend on the reweighting.

Cohort B contains more true findings: 2,554 against cohort A's 1,852, both counted before detection (the 1,857 of Table 5 are cohort A's observed findings after it). On cohort B, the toolkit detects violations of the externally testable controls at micro-F_1 0.907, with precision 0.869 and recall 0.949, against 0.909, 0.892 and 0.925 on cohort A. It flags overstated declarations at F_1 0.888, with precision 0.834 and recall 0.951, against 0.898, 0.853 and 0.947. Violation detection therefore differs from cohort A by less than the seed-to-seed variation, and contradiction detection by less than two points, so the result is not an artefact of the first generator's structure. Because cohort B shares the sensing model and the mapping, it tests robustness to population structure rather than to the detection assumptions.

Table 6: Evaluation pitfalls from Arp et al. [87] and the corresponding safeguards.

Pitfall	Guard
Data leakage	The ground truth is held separately from the toolkit input; a test scans the serialised input for any ground-truth field.
Spurious baseline	The predict-all baseline and the single-pick ceiling are reported alongside every violation-detection number.
Sampling bias	Firm demographics are drawn from national statistics; declared parameters are stated as such.
Instability	Every value can be regenerated from a fixed seed; the violation-detection F_1 is reported across thirty seeds.
Underpowered strata	Per-stratum counts are reported; strata below five firms are marked rather than read as claims.

4.7. Sample toolkit output

Algorithm 1 generates eval-000, a small firm in food and accommodation with a moderate posture and three subdomains under the apex `sme000-co.sa`: `cdn`, `mail` and `api`. The generator injects nine true external issues. The `cdn` host has an absent web application firewall, an exposed internal port and an exposed management port; the `mail` host has an exposed login panel; the `api` host has a legacy TLS version, a weak cipher, an expired certificate and end-of-life software; and the apex is missing its DNSSEC record.

The external scan observes eight of the nine, missing the end-of-life software on `api`, and produces three false positives: a firewall fingerprint on `mail` and on `api`, and an outdated software banner on `cdn`. The classification maps the resulting eleven findings to nine controls.

Figure 7 reports the two scores. The assessed score is 0.44 across the 16 externally testable subcontrols. The documented score is 0.69 across 39 controls: the 16 plus a further 23 rows supported by off-surface internal assessments or declarations. Neither score is the compliance percentage the regulator computes over the 203-row catalogue: the firm still lacks evidence for 157 of the 180 attestation-required rows, and any report containing these scores must say so. The two scores differ for every firm in the cohort. The median difference is 0.20, with a range from −0.01 to 0.40, the negative extreme being a firm whose internal assessment recorded more gaps than its external surface reveals.

The firm declared eight externally testable controls as compliant, although it actually violates four of them. The toolkit flags three of them: identity and access management (2-2-3-1), the DNS record set (2-5-3-7) and cryptography (2-8-3-3). Each is contradicted by a Tier 1 finding. The fourth overstatement, end-of-life software (2-3-3-3), survives because the scan missed the finding that would contradict it. This example illustrates the recall limit of Section 4.4 for one firm: the toolkit flags what the

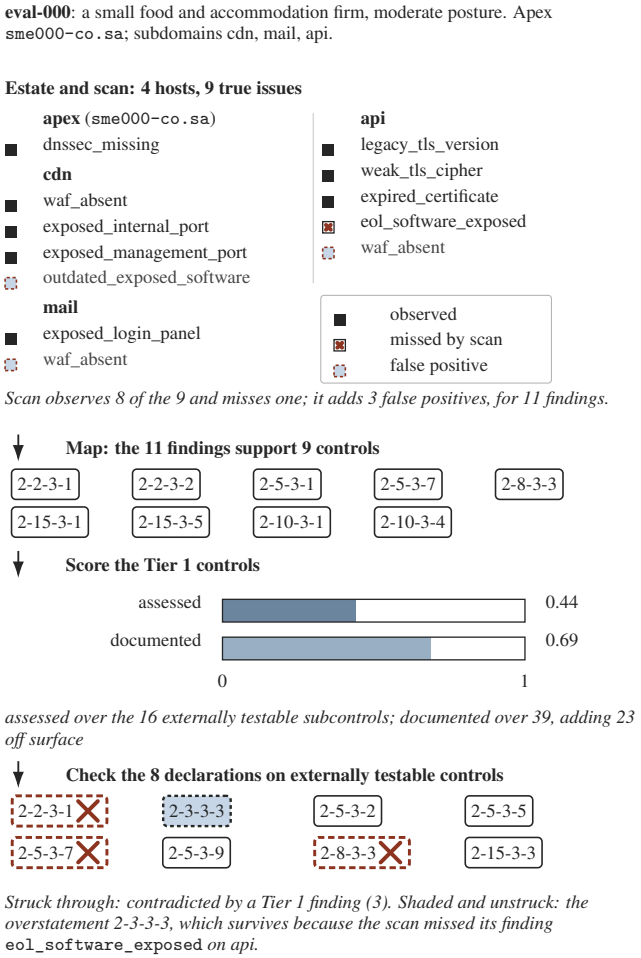


Figure 7: One firm, eval-000, traced from estate to evidence. The scan maps eleven findings to nine controls, reports each score over its corresponding set of controls, and flags three of the firm's eight declarations on externally testable controls as contradicted; a fourth overstatement (2-3-3-3) survives because the scan missed its finding.

scan observes, and a violation missed by the scan cannot trigger a contradiction flag. The firm's remaining declarations, and all of its internal assessment records, concern off-surface controls that no finding can assess.

5. Discussion

5.1. Sensing is modelled rather than measured

The violation-detection and contradiction results are integrity checks computed under modelled detection rates. Of the 31 detector-performance rates, 15 are deterministic reads, 7 are high-reliability probes, and 9 are analyst judgements bounded by published figures (Appendix A). Eight of those nine analyst-judgement rates belong to detectors with no live inspector, DKIM being the exception, so their sensitivity and false-positive rates cannot be measured on a testbed until those inspectors are implemented. The next step is a field deployment with consenting organisations, which would replace modelled rates with measured performance.

5.2. Implementation status of the detector suite

Fifteen of the 31 detectors have a live inspector and 16 are specified but not implemented. The specified detectors carry part of the violation-detection result, so implementing the remaining inspectors would allow the result to measure the implemented toolkit rather than the full detector specification. We report the implemented-only result alongside the full one to distinguish implementation coverage from specification-level evaluation.

5.3. Contradiction detection under a non-adaptive adversary

Section 3.3.6 defines the adversary as a non-adaptive declarer, so the reported F_1 of 0.898 does not describe an organisation adapting its external surface to the scan. The check also produces false-positive flags: at precision 0.853, roughly one flag in seven is a false positive, so an honest organisation can be marked as overstating a declaration it in fact meets. A contradiction flag is therefore a prompt for human review rather than a verdict.

5.4. The evaluation cohort is synthetic

No public measurement of Saudi small-firm posture exists, so we construct the cohort instead. Its demographic parameters come from national statistics, and regional incident reports inform the relative pattern of selected exposure categories (Section 3.4.1); its exposure magnitudes are declared assumptions and are not field-validated. The evaluation therefore shows how the toolkit performs on a synthetic population constructed from stated assumptions, with no field measurement anchoring it.

We considered a larger cohort but selected 100 firms: at that size the resampling interval already matches the width of the modelling interval, so a larger sample would tighten only the resampling interval and leave the dominant uncertainty, the modelled detection rates, in place. A cohort of 100 firms also represents a plausible assessment batch. The trade-off is that some sector strata are small: two of the eight sectors contain only four firms each, and their per-sector figures are underpowered, so we report them with their counts.

The toolkit receives the generated estate rather than discovering it, so asset-discovery recall is assumed rather than tested. A proper test of enumerator recall by domain size requires a certificate transparency ground truth over many real domains; a future field study must establish it.

5.5. Concentration of detectors on a single control

Ten of the 31 detectors map to control 2-15-3-3, so one interpretation of one web application control determines much of the externally testable coverage. That reading is contestable: under a coarser reading, seven of those ten detectors map to the parent control 2-15-3 under the control's *as a minimum* clause. Table 7 reports the principal figures under both readings, with the relabelling applied to predicted and gold labels alike.

The coarse reading mainly changes the concentration measures: the finding-weighted modal share falls from 0.348 to 0.236, and the degenerate constant-mapper baseline from 0.319 to 0.216. The main performance results change little and

Table 7: The principal figures under the fine reading of control 2-15-3-3 (the released reading, a leaf) and the coarse reading, which reassigns its seven hardening detectors to the parent 2-15-3. The relabelling is applied to predicted and gold labels alike. The coarse reading changes the concentration measures most and the violation-detection and contradiction figures least.

Quantity	Fine	Coarse
Violation-detection precision	0.892	0.899
Violation-detection recall	0.925	0.930
Violation-detection F_1	0.909	0.915
Contradiction F_1	0.898	0.921
Modal share (weighted)	0.348	0.236
Constant-mapper baseline F_1	0.319	0.216

increase slightly: the violation-detection F_1 rises from 0.909 to 0.915 and the contradiction F_1 from 0.898 to 0.921, because mapping a finding to the parent causes it to contradict a broader set of sibling declarations. The externally testable row count does not change, since the leaf retains its three protocol detectors and the parent is already one of the seven counted parents.

We retain the fine reading as primary because the released classification is the object being evaluated. Figure 2 shows the concentration directly.

5.6. Status of the curation

The tiers are author-curated against the standard's text. On review, 42 of the 200 ECC-2:2024 rows received a provisional default, and review confirmed every assignment against the control text; the remainder were assigned directly from explicit wording in the control text. They have not been validated by an independent expert panel. The validation instrument has been prepared but not fielded. It comprises an analysis plan specified before fielding and a set of per-expert materials. The paper therefore reports no expert-agreement figure because the study has not been run, not because validation was judged unnecessary. Independent expert validation of the externally testable tier is the main outstanding requirement and remains future work, consistent with the finding that self-assessment in this setting requires independent checks [13].

As a preliminary check before independent panel validation, one open-weight model (Qwen2.5 7B) read each control's own text and classified it on the primary binary boundary (externally testable or attestation-required) without access to our classification. It agrees on 187 of the 203 rows, a raw agreement of 0.92, at a Cohen's κ of 0.47; the raw figure is inflated by the attestation-required class that dominates the catalogue, so the moderate κ is the more informative summary. The two readings differ on 16 rows. Fifteen are controls we mark externally testable that the model classifies as attestation-required. Of those fifteen, seven are parent controls classified as Tier 1 because an observable child sits beneath them; the remaining eight are transport security, open port and web application exposure rows. The sixteenth is an awareness control (1-10-3-3) the model classifies as externally testable; awareness leaves no external trace, so we retain it as a declaration row. We release the disagreement set with the artefacts so that independent reviewers can judge the 16 rows. Our own re-reading of the control text left every assignment in place, but that re-reading is not independent.

This model check, performed without access to our labels, uses a much coarser task than the 31-detector mapping at which the baselines fail. It is a limited stability check of the boundary, not evidence that the classification is correct. Materials for a blind intra-coder recode of all 203 rows by the author after an interval have been prepared, but the recode has not yet been conducted.

5.7. Implications

The one-in-nine share has different implications for each audience. For an SME, the externally testable tier is the subset of the catalogue on which one authorised scan can supply independent evidence of observable violations, and the contradiction check lets the same scan check the organisation's own declarations before any assessor is engaged. For an assessor, the classification separates the rows for which a scan can supply violation evidence from the rows that require records or interviews, so assessment effort can focus on the controls external observation cannot test. For the regulator, a channel that relies on organisation-submitted declarations leaves even the observable subset unverified; attaching external observation to an existing channel would support national-scale checks for observable violations on the 23 externally testable rows. And for compliance platforms, the classification gives an external-evidence module a defined scope: the 23 rows it can assess, and the 180 that require organisation-supplied evidence.

6. Conclusion

Unauthenticated external observation can test 16 of the 92 ECC-2:2024 subcontrols. When their seven parent rows are counted, this corresponds to 23 of the 203 combined catalogue rows (11.3%), approximately one in nine. The classification of every control into externally testable, internal assessment and declaration is the paper's central contribution. The released mapping, toolkit and cohort support independent reproduction of every reported value from a single seed. On the synthetic cohort the toolkit detects violations of the externally testable controls well above the predict-all baseline and flags contradicted declarations on the controls the scan observes. Both results were obtained under a parameterised sensing model and an author-curated classification; independent expert validation of that classification is the next step. The observation boundary is not specific to the Saudi control sets, and the same classification could be applied to another national or international framework. The practical value is limited but clear. For an organisation with no security staff, the toolkit identifies which controls an external scan can test and where internal evidence or professional assessment must take over.

References

[1] International Organization for Standardization, International Electrotechnical Commission, Information security, cybersecurity and privacy protection — Information security management systems — Requirements, International Standard ISO/IEC 27001:2022, ISO/IEC, Geneva, Switzerland (Oct. 2022).

[2] National Institute of Standards and Technology, The NIST cybersecurity framework (CSF) 2.0, NIST Cybersecurity White Paper NIST CSWP 29, National Institute of Standards and Technology, Gaithersburg, MD (Feb. 2024). doi:10.6028/NIST.CSWP.29.

[3] Center for Internet Security, CIS critical security controls version 8, Tech. rep., Center for Internet Security, East Greenbush, NY (2021).

[4] N. Keller, M. Barrett, S. Quinn, K. Scarfone, M. C. Smith, V. Johnson, National Online Informative References (OLIR) Program: Overview, Benefits, and Use, Tech. Rep. NIST IR 8278 Revision 1, National Institute of Standards and Technology (2024). doi:10.6028/NIST.IR.8278r1.

[5] F. Angermeir, J. Fischbach, F. Moyón, D. Méndez, Towards Automated Continuous Security Compliance, in: Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '24), ACM, Barcelona, Spain, 2024. doi:10.1145/3674805.3690748.

[6] C. Martínez, I. Etxaniz, A. Molinuevo, J. Alonso, MEDINA Catalogue of Cloud Security Controls and Metrics: Towards Continuous Cloud Security Compliance, Open Research Europe 4 (2024) 90. doi:10.12688/openreseurope.16669.1.

[7] T. R. McIntosh, T. Sušnjak, T. Liu, P. Watters, D. Xu, D. Liu, R. Nowrozy, M. N. Halgamuge, From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models, Computers & Security 144 (2024) 103964. doi:10.1016/j.cose.2024.103964.

[8] O. Briliyant, A. Javed, Y. Cherdantseva, Beyond the automation gap: A survey of continuous audit for IoT security, Computers & Security 168 (2026) 104953. doi:10.1016/j.cose.2026.104953.

[9] A. Chidukwani, S. Zander, P. Koutsakis, A Survey on the Cyber Security of Small-to-Medium Businesses: Challenges, Research Focus and Recommendations, IEEE Access 10 (2022) 85701–85719. doi:10.1109/ACCESS.2022.3197899.

[10] A. Chidukwani, S. Zander, P. Koutsakis, Cybersecurity Preparedness of Small-to-Medium Businesses: A Western Australia Study with Broader Implications, Computers & Security 145 (2024) 104026. doi:10.1016/j.cose.2024.104026.

[11] N. Huaman, B. von Skarczinski, C. Stransky, D. Wermke, Y. Acar, A. Dreißigacker, S. Fahl, A large-scale interview study on information security in and attacks against small and medium-sized enterprises, in: 30th USENIX Security Symposium (USENIX Security 21), USENIX Association, 2021, pp. 1235–1252.

[12] A. Cartwright, E. Cartwright, Underinvestment in Cyber Security: Quantifying Cyber Security Behavior in UK Businesses, Journal of Small Business Management (2025). doi:10.1080/00472778.2025.2549068.

[13] A. Chidukwani, S. Zander, P. Koutsakis, Beyond self-reporting: Uncovering the operational realities of SME cybersecurity through expert assessment, Computers & Security 164 (2026) 104839. doi:10.1016/j.cose.2026.104839.

[14] A. Alfaadhel, I. Almomani, M. Ahmed, Risk-based cybersecurity compliance assessment system (RC2AS), Applied Sciences 13 (10) (2023) 6145. doi:10.3390/app13106145.

[15] I. Almomani, M. Ahmed, L. Maglaras, Cybersecurity maturity assessment framework for higher education institutions in Saudi Arabia, PeerJ Computer Science 7 (2021) e703. doi:10.7717/peerj-cs.703.

[16] I. Almomani, S. Mehdawi, Y. Allabadi, Towards effective cybersecurity governance: Jordan compliance system and self-assessment tools, Journal of Cybersecurity and Privacy 6 (2) (2026) 60. doi:10.3390/jcp6020060.

[17] A. Palma, G. Acitelli, A. Marrella, S. Bonomi, M. Angelini, A compliance assessment system for incident management process, Computers & Security 146 (2024) 104070. doi:10.1016/j.cose.2024.104070.

[18] N. Mexis, S. Katzenbeisser, A data-driven approach for cyber security assessments of SMEs, in: Availability, Reliability and Security (ARES 2025), Vol. 15997 of Lecture Notes in Computer Science, Springer, Cham, 2025. doi:10.1007/978-3-032-00639-4_2.

[19] S. M. Ali, A. Razzaque, M. Yousaf, R. U. Shan, An automated compliance framework for critical infrastructure security through artificial intelligence, IEEE Access 13 (2024) 4436–4459. doi:10.1109/access.2024.3524496.

[20] A. B. A. Ali, R. K. Ayyasamy, R. Akbar, A. K. Jebna, K. Adnan, Cybersecurity infrastructure compliance key factors to detect and mitigate malware attacks in SMEs: A systematic literature review, SAGE Open 15 (2025). doi:10.1177/21582440251314671.

[21] National Institute of Standards and Technology, Open Security Controls Assessment Language (OSCAL), accessed: 2026-05-18 (2026). URL https://pages.nist.gov/OSCAL/

[22] D. Waltermire, S. Quinn, H. Booth, K. Scarfone, D. Prisaca, The Technical Specification for the Security Content Automation Protocol (SCAP): SCAP Version 1.3, Tech. Rep. NIST Special Publication 800-126 Revision 3, National Institute of Standards and Technology (2018). doi:10.6028/NIST.SP.800-126r3.

[23] S. Fenz, T. Neubauer, Ontology-based information security compliance determination and control selection on the example of ISO 27002, Information & Computer Security 26 (5) (2018) 551–567. doi:10.1108/ICS-02-2018-0020.

[24] Z. Durumeric, E. Wustrow, J. A. Halderman, ZMap: Fast internet-wide scanning and its security applications, in: 22nd USENIX Security Symposium (USENIX Security 13), USENIX Association, Washington, D.C., 2013, pp. 605–620.

[25] Z. Durumeric, D. Adrian, A. Mirian, M. Bailey, J. A. Halderman, A search engine backed by Internet-wide scanning, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2015, pp. 542–553. doi:10.1145/2810103.2813703.

[26] Z. Durumeric, D. Adrian, A. Mirian, J. Kasten, E. Bursztein, N. Lidzborski, K. Thomas, V. Eranti, M. Bailey, J. A. Halderman, Neither snow nor rain nor MITM...: An empirical analysis of email delivery security, in: Proceedings of the 2015 ACM Internet Measurement Conference (IMC), ACM, 2015, pp. 27–39. doi:10.1145/2815675.2815695.

[27] C. Wang, K. Shen, M. Guo, Y. Zhao, M. Zhang, J. Chen, B. Liu, X. Zheng, H. Duan, Y. Lin, Q. Pan, A large-scale and longitudinal measurement study of DKIM deployment, in: 31st USENIX Security Symposium (USENIX Security 22), USENIX Association, 2022, pp. 1185–1201, distinguished Paper Award; USENIX assigns no DOI.

[28] P. K. Manadhata, J. M. Wing, An Attack Surface Metric, IEEE Transactions on Software Engineering 37 (3) (2011) 371–386. doi:10.1109/TSE.2010.60.

[29] M. Vermeer, J. West, A. Cuevas, S. Niu, N. Christin, M. van Eeten, T. Fiebig, C. Gañán, T. Moore, SoK: A framework for asset discovery: Systematizing advances in network measurements for protecting organizations, in: 2021 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2021, pp. 440–456. doi:10.1109/EuroSP51992.2021.00037.

[30] N. Gelernter, H. Schulmann, M. Waidner, External attack-surface of modern organizations, in: Proceedings of the 19th ACM Asia Conference on Computer and Communications Security (ASIA CCS '24), Association for Computing Machinery, 2024, pp. 589–604. doi:10.1145/3634737.3656295.

[31] M. Husák, L. Sadlek, Attack Surface Management: State of the Art and Operational Challenges, in: 2025 IEEE 11th International Conference on Network Softwarization (NetSoft), IEEE, Budapest, Hungary, 2025. doi:10.1109/NetSoft64993.2025.11080588.

[32] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, M. Bailey, M. Liu, Cloudy with a chance of breach: Forecasting cyber security incidents, in: 24th USENIX Security Symposium (USENIX Security 15), USENIX Association, Washington, D.C., 2015, pp. 1009–1024. URL https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/liu

[33] A. Sarabi, P. Naghizadeh, Y. Liu, M. Liu, Risky business: Fine-grained data breach prediction using business profiles, Journal of Cybersecurity 2 (1) (2016) 15–28. doi:10.1093/cybsec/tyw004.

[34] H. Hammouchi, N. Nejjari, G. Mezzour, M. Ghogho, H. Benbrahim, STRisk: A socio-technical approach to assess hacking breaches risk, IEEE Transactions on Dependable and Secure Computing 20 (2) (2023) 1074–1087. doi:10.1109/TDSC.2022.3149208.

[35] S. J. Choi, M. E. Johnson, The relationship between cybersecurity ratings and the risk of hospital data breaches, Journal of the American Medical Informatics Association 28 (10) (2021) 2085–2092. doi:10.1093/jamia/ocab142.

[36] D. W. Woods, R. Böhme, SoK: Quantifying cyber risk, in: 2021 IEEE Symposium on Security and Privacy (S&P), IEEE, 2021, pp. 211–228. doi:10.1109/SP40001.2021.00053.

[37] M. Safaei Pour, C. Nader, K. Friday, E. Bou-Harb, A comprehensive survey of recent Internet measurement techniques for cyber security, Computers & Security 128 (2023) 103123. doi:10.1016/j.cose.2023.103123.

[38] C. Bennett, A. Abdou, P. C. van Oorschot, Empirical Scanning Analysis of Censys and Shodan, in: Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb), Network and Distributed System Security Symposium (NDSS), Internet Society, 2021. doi:10.14722/madweb.2021.23009.

[39] O. F. Keskin, K. M. Caramancion, I. Tatar, O. Raza, U. Tatar, Cyber third-party risk management: A comparison of non-intrusive risk scoring reports, Electronics 10 (10) (2021) 1168. doi:10.3390/electronics10101168.

[40] S. Rahaman, G. Wang, D. Yao, Security certification in payment card industry: Testbeds, measurements, and recommendations, in: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2019, pp. 481–498. doi:10.1145/3319535.3363195.

[41] United Kingdom National Cyber Security Centre, Cyber Essentials: Overview and Requirements, accessed: 2026-07-28 (2026).
URL https://www.ncsc.gov.uk/cyberessentials/overview

[42] National Cybersecurity Authority, Essential Cybersecurity Controls (ECC-2:2024), Official regulatory document, accessed: 2026-05-18 (2024).
URL https://cdn.nca.gov.sa/api/files/public/upload/86e09090-44e4-481f-bc28-355673607654_ECC--2024-EN.pdf

[43] National Cybersecurity Authority, Non-CNI Private Sector Entities Cybersecurity Controls (NCNICC-1:2025), Official regulatory document, accessed: 2026-05-18 (2025).
URL https://cdn.nca.gov.sa/api/files/public/upload/721bed5c-5304-497c-a5ac-a92a049e6306_NCNICC-AR-.pdf

[44] National Cybersecurity Authority, Essential Cybersecurity Controls (ECC-1:2018), Official regulatory document, referenced only SPF for email domain validation under subcontrol 2-4-3-5; 5 domains, 29 subdomains, 114 controls (2018).

[45] Joint Task Force, Security and Privacy Controls for Information Systems and Organizations, Tech. Rep. NIST Special Publication 800-53 Revision 5, National Institute of Standards and Technology (2020). doi:10.6028/NIST.SP.800-53r5.

[46] Saudi Arabian Monetary Authority, Cyber Security Framework, Version 1.0 (May 2017).
URL https://www.sama.gov.sa

[47] Kingdom of Saudi Arabia, Vision 2030 Annual Report 2025, Vision 2030 Program, accessed: 2026-07-07 (2026).
URL https://www.vision2030.gov.sa/media/ecdjfopq/vision2030_annual_report_2025_en.pdf

[48] Kingdom of Saudi Arabia, Personal Data Protection Law, Royal Decree M/19, article 29 governs cross-border personal data transfer (2021).

[49] National Cybersecurity Authority, NCA Launches the Cybersecurity Enablement for Small and Medium Enterprises, NCA news release, accessed: 2026-05-18 (Nov. 2025).
URL https://nca.gov.sa/en/news/2050/

[50] National Data Management Office, National Data Governance Interim Regulations, Saudi Data and AI Authority (SDAIA) (2020).

[51] Ministry of Communications and Information Technology, KSA Cloud First Policy, accessed: 2026-05-19 (Oct. 2020).
URL https://www.mcit.gov.sa/sites/default/files/cloud_policy_en.pdf

[52] C. Bronk, E. Tikk-Ringas, The cyber attack on Saudi Aramco, Survival 55 (2) (2013) 81–96. doi:10.1080/00396338.2013.784468.

[53] S. Zhioua, The Middle East under malware attack: Dissecting cyber weapons, in: 2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops, IEEE, 2013, pp. 11–16. doi:10.1109/ICDCSW.2013.30.

[54] Small and Medium Enterprises General Authority (Monsha'at), SME Monitor Q2 2025, Official report, accessed: 2026-05-18 (2025).
URL https://www.monshaat.gov.sa/sites/default/files/2025-09/V5.0%20Monsha%27at%20SMEM%20Report%20-%20Q2-25.pdf

[55] Saudi Ministry of Commerce, Ministry of Commerce Issues Q2 2025 Business Sector Bulletin, accessed: 2026-05-18 (Jul. 2025).
URL https://mc.gov.sa/en/mediacenter/News/Pages/06-07-25-01.aspx

[56] National Cybersecurity Authority, Key Economic Indicators in the Cybersecurity Sector in the Kingdom 2025, accessed: 2026-05-19 (2025).
URL https://nca.gov.sa/en/studies-and-reports/reports/key-economic-indicators-in-the-cybersecurity-sector/

[57] National Cybersecurity Authority, Haseen: National Portal for Cybersecurity Services, accessed: 2026-07-07 (2026).
URL https://haseen.gov.sa/

[58] National Cybersecurity Authority, Verification of Email Authentication Protocol Implementation, NCA service page, accessed: 2026-05-18 (2025).
URL https://nca.gov.sa/en/services/verification-of-email-authentication-protocol-implementation/

[59] D. Liu, S. Hao, H. Wang, All Your DNS Records Point to Us: Understanding the Security Threats of Dangling DNS Records, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), ACM, 2016, pp. 1414–1425. doi:10.1145/2976749.2978387.

[60] United States Department of Homeland Security, Binding Operational Directive 18-01: Enhance Email and Web Security, accessed: 2026-07-28 (2017).
URL https://www.cisa.gov/news-events/directives/bod-18-01-enhance-email-and-web-security

[61] A. Alzubaidi, Measuring the Level of Cyber-Security Awareness for Cybercrime in Saudi Arabia, Heliyon 7 (1) (2021) e06016. doi:10.1016/j.heliyon.2021.e06016.

[62] F. Alharbi, M. H. Alsulami, A. AL-Solami, Y. S. Al-Otaibi, M. Al-Osimi, F. Al-Qanor, K. Al-Otaibi, The Impact of Cybersecurity Practices on Cyberattack Damage: The Perspective of Small Enterprises in Saudi Arabia, Sensors 21 (20) (2021) 6901. doi:10.3390/s21206901.

[63] N. Rawindaran, L. Nawaf, S. Alarifi, D. Alghazzawi, F. Carroll, I. Katib, C. Hewage, Enhancing cyber security governance and policy for SMEs in Industry 5.0: A comparative study between Saudi Arabia and the United Kingdom, Digital 3 (3) (2023) 200–231. doi:10.3390/digital3030014.

[64] J. Zhang, Z. Durumeric, M. Bailey, M. Liu, M. Karir, On the mismanagement and maliciousness of networks, in: Proceedings of the 2014 Network and Distributed System Security Symposium (NDSS), Internet Society, 2014. doi:10.14722/ndss.2014.23057.

[65] D. Woods, I. Agrafiotis, J. R. C. Nurse, S. Creese, Mapping the coverage of security controls in cyber insurance proposal forms, Journal of Internet Services and Applications 8 (1) (2017) 8. doi:10.1186/s13174-017-0059-y.

[66] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, N. Vallina-Rodríguez, A long way to the top: Significance, structure, and stability of Internet top lists, in: Proceedings of the 2018 ACM Internet Measurement Conference (IMC), ACM, 2018, pp. 478–493. doi:10.1145/3278532.3278574.

[67] A. Erden, Modern web application security in practice: National-scale measurement and managerial insights from Türkiye, Computers & Security 164 (2026) 104847. doi:10.1016/j.cose.2026.104847.

[68] A. S. Alhazmi, A. Daghistani, Privacy practices of popular websites in Saudi Arabia, Journal of Umm Al-Qura University for Engineering and Architecture 16 (1) (2024) 19–29. doi:10.1007/s43995-024-00085-x.

[69] B. Edwards, J. Jacobs, S. Forrest, Risky business: Assessing security with external measurements, arXiv preprint (2019). arXiv:1904.11052, doi:10.48550/arXiv.1904.11052.

[70] ProjectDiscovery, Nuclei: Template-based vulnerability scanner, accessed: 2026-07-01 (2026).
URL https://github.com/projectdiscovery/nuclei

[71] EnableSecurity, wafw00f: Web Application Firewall Fingerprinting Tool, accessed: 2026-07-04 (2026).

URL https://github.com/EnableSecurity/wafw00f

[72] Dutch Internet Standards Platform, Internet.nl: Test for Modern Internet Standards, accessed: 2026-07-28 (2026).
URL https://internet.nl/about/

[73] S. Singanamalla, E. H. B. Jang, R. Anderson, T. Kohno, K. Heimerl, Accept the risk and continue: Measuring the long tail of government HTTPS adoption, in: Proceedings of the 2020 ACM Internet Measurement Conference (IMC), ACM, 2020, pp. 577–597. doi:10.1145/3419394.3423645.

[74] S. Kitterman, Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1, RFC 7208, Internet Engineering Task Force (Apr. 2014).
URL https://www.rfc-editor.org/info/rfc7208

[75] D. Crocker, T. Hansen, M. Kucherawy, DomainKeys Identified Mail (DKIM) Signatures, RFC 6376, Internet Engineering Task Force (Sep. 2011).
URL https://www.rfc-editor.org/info/rfc6376

[76] M. Kucherawy, E. Zwicky, Domain-based Message Authentication, Reporting, and Conformance (DMARC), RFC 7489, Internet Engineering Task Force (Mar. 2015).
URL https://www.rfc-editor.org/info/rfc7489

[77] Forum of Incident Response and Security Teams, Common Vulnerability Scoring System version 3.1: Specification Document (2019).
URL https://www.first.org/cvss/v3.1/specification-document

[78] Small and Medium Enterprises General Authority (Monsha'at), SME monitor, fourth quarter 2023: enterprise size distribution, https://www.monshaat.gov.sa/en/node/53859, accessed: 2026-08-09 (2024).

[79] Small and Medium Enterprises General Authority (Monsha'at), SME definition, https://www.monshaat.gov.sa/en/SMEs-definition, accessed: 2026-08-09 (2024).

[80] Small and Medium Enterprises General Authority (Monsha'at), SME monitor, first quarter 2022: sector and regional distribution, https://monshaat.gov.sa/sites/default/files/2022-06/Monshaat%20Q1%202022%20-.pdf, accessed: 2026-08-09 (2022).

[81] General Authority for Statistics (GASTAT), Statistics of small and medium enterprises 2023 (comprehensive economic survey), https://www.stats.gov.sa/documents/20117/2435267/SEM2023EN+(1).pdf, accessed: 2026-08-09 (2024).

[82] General Authority for Statistics (GASTAT), Establishments' ICT access and usage statistics 2024, https://www.stats.gov.sa/documents/20117/2435267/Establishments+ICT+Access+and+Usage+Statistics+2024-EN.pdf, accessed: 2026-08-09 (2025).

[83] Communications, Space and Technology Commission (CST), Saudi domain name report 2024, https://www.cst.gov.sa/en/mediacenter/pressreleases/Pages/2024051201.aspx, accessed: 2026-08-09 (2024).

[84] sirar by stc, Threat landscape report in 2023, https://www.sirar.com.sa/wp-content/uploads/2024/02/Threat-Landscape-Report-in-2023-Final.pdf, vendor threat report used only to justify generator priors; archived at https://web.archive.org/web/20241125021337/https://www.sirar.com.sa/wp-content/uploads/2024/02/Threat-Landscape-Report-in-2023-Final.pdf; accessed 2026-08-16 (2024).

[85] Verizon, 2025 data breach investigations report, https://www.verizon.com/business/resources/reports/2025-dbir-data-breach-investigations-report.pdf, vendor incident report used only to justify generator priors; archived at https://web.archive.org/web/20260812112850/https://www.verizon.com/business/resources/reports/2025-dbir-data-breach-investigations-report.pdf; accessed 2026-08-16 (2025).

[86] SOCRadar, Saudi arabia threat landscape report 2024, https://socradar.io/wp-content/uploads/2024/12/SOCRadar-Saudi-Arabia-Threat-Landscape-Report-2024.pdf, vendor threat report used only to justify generator priors; archived at https://web.archive.org/web/20250419013746/https://socradar.io/wp-content/uploads/2024/12/SOCRadar-Saudi-Arabia-Threat-Landscape-Report-2024.pdf; accessed 2026-08-16 (2024).

[87] D. Arp, E. Quiring, F. Pendlebury, A. Warnecke, F. Pierazzi, C. Wressnegger, L. Cavallaro, K. Rieck, Dos and don'ts of machine learning in computer security, in: 31st USENIX Security Symposium (USENIX Security 22), USENIX Association, 2022, pp. 3971–3988.
URL https://www.usenix.org/conference/usenixsecurity22/presentation/arp

[88] HTTP Archive, The 2024 Web Almanac: Security, https://almanac.httparchive.org/en/2024/security, annual measurement report, no DOI; archived at https://web.archive.org/web/20260515050332/https://almanac.httparchive.org/en/2024/security; accessed 2026-08-12 (2024).

[89] HTTP Archive, Security feature adoption by country, 2024 Web Almanac queries, https://github.com/HTTPArchive/almanac.httparchive.org/blob/main/sql/2024/security/feature_adoption_by_country.sql, per-country slice of the same crawl the security chapter reports (2024-06-01; country assignment from the Chrome UX Report); the Saudi figures are reproducible from the published SQL over the public BigQuery dataset. Accessed 2026-08-16 (2024).

[90] SaudiNIC, Saudi network information center registry statistics, https://nic.sa/en/statistics, official registry counts: 93,154 registered national domains, 2,180 of them DNSSEC-signed. Accessed 2026-08-16 (2026).

[91] Qwen Team, Qwen2.5 Technical Report, arXiv preprint arXiv:2412.15115, model card for the Qwen2.5-7B evaluated model. Accessed: 2026-08-09 (2024). arXiv:2412.15115, doi:10.48550/arXiv.2412.15115.

[92] Gemma Team, Google DeepMind, Gemma 3 Technical Report, arXiv preprint arXiv:2503.19786, model card for the Gemma-3-12B evaluated model. Accessed: 2026-08-09 (2025). arXiv:2503.19786, doi:10.48550/arXiv.2503.19786.

[93] M. Abdin, et al., Phi-4 Technical Report, Microsoft Research; arXiv preprint arXiv:2412.08905, model card for the Phi-4 (14B) evaluated model. Accessed: 2026-08-09 (2024). arXiv:2412.08905, doi:10.48550/arXiv.2412.08905.

[94] Qwen Team, Qwen3 Technical Report, arXiv preprint arXiv:2505.09388, model card for the Qwen3-14B and Qwen3-32B evaluated models. Accessed: 2026-08-09 (2025). arXiv:2505.09388, doi:10.48550/arXiv.2505.09388.

[95] DeepSeek-AI, DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv preprint arXiv:2501.12948, model card for the DeepSeek-R1-Distill-Qwen-14B evaluated model. Accessed: 2026-08-09 (2025). arXiv:2501.12948, doi:10.48550/arXiv.2501.12948.

[96] OpenAI, gpt-oss-120b and gpt-oss-20b Model Card, https://openai.com/index/introducing-gpt-oss/, model card for the gpt-oss-20b evaluated model. Accessed: 2026-08-09 (2025). arXiv:2508.10925.

[97] Mistral AI, Mistral Small 3, Model card, https://huggingface.co/mistralai/Mistral-Small-24B-Instruct-2501, model card for the Mistral-Small-24B evaluated model. Accessed: 2026-08-09 (2025).
URL https://mistral.ai/news/mistral-small-3/

[98] M. S. Bari, Y. Alnumay, N. A. Alzahrani, N. M. Alotaibi, H. A. Alyahya, S. AlRashed, et al., ALLaM: Large Language Models for Arabic and English, arXiv preprint arXiv:2407.15390 (2024). arXiv:2407.15390, doi:10.48550/arXiv.2407.15390.

[99] Cohere, Command R7B Arabic, Model card, https://huggingface.co/CohereLabs/c4ai-command-r7b-arabic-02-2025, model card for the command-r7b-arabic (7B) evaluated model. Accessed: 2026-08-09 (2025).
URL https://cohere.com/blog/command-r7b-arabic

[100] N. Sengupta, S. K. Sahu, B. Jia, S. Katipomu, H. Li, F. Koto, et al., Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models, arXiv preprint arXiv:2308.16149 (2023). arXiv:2308.16149, doi:10.48550/arXiv.2308.16149.

[101] SILMA AI, SILMA-9B-Instruct-v1.0, Model card, https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0, model card for the SILMA-9B-Instruct evaluated model. Accessed: 2026-08-09 (2024).

[102] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, in: Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 2318–2335. doi:10.18653/v1/2024.findings-acl.137.

Appendix A. Modelled detection rates and generator priors

Tables A.8, A.9 and A.10 give the modelled detection rates, the generator priors, and the incident-informed exposure-weight columns described in Section 3.4. All three are generated directly from the same constants that construct the cohort, so the reported values remain consistent with the generated cohort.

Table A.8: Modelled detection rates for the 31 detectors, with their provenance. Sensitivity is the probability that the scanner confirms a present issue on a reachable host; FA is the false-positive rate, the probability that it reports an absent one. A deterministic read retrieves a published value and has no modelled false-positive rate; a high-reliability probe is reliable (at least 0.95) but raises occasional false positives; analyst judgement is used for ambiguous fingerprint, banner, or path-limited signals. These are abbreviated below as read, probe and judgement. The Live column identifies detectors with an implemented inspector.

Detector	Sens.	FA	Provenance	Live
`missing_hsts`	0.99	0.00	read	yes
`missing_csp`	0.99	0.00	read	yes
`missing_x_frame_options`	0.99	0.00	read	yes
`missing_x_content_type_options`	0.99	0.00	read	yes
`missing_referrer_policy`	0.99	0.00	read	yes
`no_https_redirect`	0.98	0.00	read	yes
`missing_permissions_policy`	0.99	0.00	read	no
`insecure_cookie`	0.98	0.00	read	no
`expired_certificate`	0.99	0.00	read	yes
`self_signed_certificate`	0.98	0.00	read	no
`weak_certificate_key`	0.97	0.00	read	no
`spf_missing`	0.99	0.00	read	yes
`dmarc_missing_or_none`	0.99	0.00	read	yes
`mta_sts_missing`	0.98	0.00	read	no
`dnssec_missing`	0.99	0.00	read	no
`legacy_tls_version`	0.96	0.02	probe	yes
`weak_tls_cipher`	0.96	0.02	probe	yes
`exposed_internal_port`	0.95	0.04	probe	yes
`exposed_management_port`	0.95	0.04	probe	yes
`high_risk_port_open`	0.95	0.04	probe	yes
`exposed_database_port`	0.95	0.04	probe	no
`exposed_dev_environment`	0.96	0.04	probe	no
`directory_listing_enabled`	0.92	0.03	judgement	no
`exposed_sensitive_file`	0.75	0.06	judgement	no
`waf_absent`	0.90	0.20	judgement	no
`outdated_exposed_software`	0.72	0.06	judgement	no
`eol_software_exposed`	0.70	0.08	judgement	no
`exposed_login_panel`	0.88	0.08	judgement	no
`dkim_missing`	0.93	0.24	judgement	yes
`dns_zone_transfer_open`	0.90	0.02	judgement	no
`no_ddos_protection`	0.85	0.20	judgement	no

The declared exposure rates are conservative. For every detector with a published benchmark, the cohort's rate is below the reported non-adoption level. A content security policy is absent from 0.26 of the 365 hosts, against 0.81 globally in 2024 and 0.76 for sites with a predominantly Saudi audience in the same crawl [88, 89]. A DNSSEC record is absent for 0.54 of firms, against 0.98 of registered Saudi national domains left unsigned [90]. And SPF, DMARC and DKIM records are absent for 0.18, 0.33 and 0.21 of firms respectively, below the non-adoption levels reported by internet-scale email measurement [26, 27, 38]. The rates are therefore conservative and are not inflated to make failures easier to detect. A population with no security staff is plausibly more exposed than this, and since precision rises with the true finding rate, the reported precision is a conservative estimate under these assumptions. A sector-disaggregated measurement of Saudi SME posture requires future work.

Table A.9: Generator population priors with their sources. The firm size shares and the major sector prevalences are drawn from national statistics; everything else, including the three minor sector shares, is a declared modelling assumption, because no Saudi source measures it. Values are generated directly from the generator's constants, so the table remains consistent with the cohort it produced.

Parameter	Value	Source
Firm size bands, share of Class B		
Small, apex plus 1 to 4 subdomains	0.89	Monsha'at Q4 2023
Medium, apex plus 3 to 8 subdomains	0.11	Monsha'at Q4 2023
Security posture, share of the size band		
Strong, small firm	0.15	declared
Moderate, small firm	0.35	declared
Weak, small firm	0.50	declared
Strong, medium firm	0.30	declared
Moderate, medium firm	0.40	declared
Weak, medium firm	0.30	declared
Posture multiplier		
Strong	0.20	declared
Moderate	0.55	declared
Weak	1.00	declared
Sector prevalence		
Retail and wholesale	0.27	Monsha'at Q1 2022
Construction	0.20	Monsha'at Q1 2022
Food and accommodation	0.10	Monsha'at Q1 2022
Professional and admin services	0.14	Monsha'at Q1 2022
Manufacturing	0.10	declared
Transport and logistics	0.05	Monsha'at Q1 2022
Education	0.07	declared
Health	0.07	declared
Domain TLD weight		
`com`	0.55	CST 2024
`com.sa`	0.20	CST 2024
`sa`	0.15	CST 2024
`net.sa`	0.10	CST 2024
Evidence model, on the scanned controls and off surface		
Declare a clean control	0.70	declared
Overstate a violated control	0.50	declared
Declaration expired at the epoch	0.10	declared
Declaration not yet effective	0.05	declared
Off-surface rows with evidence, strong firm	0.20	declared
Off-surface rows with evidence, moderate firm	0.12	declared
Off-surface rows with evidence, weak firm	0.05	declared
Internal assessment finds a gap	0.25	declared

Table A.10: The four incident-informed columns of the sector-by-category exposure-weight matrix: ports, outdated software, end-of-life software and authentication surfaces. Direction and rank follow the regional incident record of Section 3.4.1; the magnitudes remain declared. The seven other categories are unchanged and stay declared. Values are generated directly from the generator's constants, so the table remains consistent with the cohort it produced.

Sector	Ports	Software	End of life	Auth
Retail and wholesale	0.50	0.56	0.48	0.54
Construction	0.55	0.62	0.62	0.50
Food and accommodation	0.48	0.58	0.52	0.50
Professional and admin services	0.50	0.55	0.50	0.56
Manufacturing	0.60	0.65	0.65	0.50
Transport and logistics	0.60	0.58	0.56	0.54
Education	0.55	0.64	0.60	0.72
Health	0.55	0.62	0.56	0.75

Appendix B. Evaluated baseline models

The mapping experiment of Section 4.2 evaluates twelve open-weight language models and one embedding model, all run zero-shot at temperature 0 under a fixed seed (Table B.11).

A local probe of Qwen2.5 7B with and without the observation boundary of Section 3.2.1 increases micro-F_1 only from 0.13 to 0.16. The main difficulty is therefore matching each finding to the relevant control text, not missing the boundary definition. A leave-one-out few-shot configuration, with the target mapping held out, achieves micro-F_1 0.00, below

Table B.11: The models scored against the curated mapping. General and Arabic-specialised language models span 7 to 32 billion parameters; the embedding model serves the semantic baseline.

Model	Parameters	Role
Qwen2.5 [91]	7B	general
Gemma 3 [92]	12B	general
Phi-4 [93]	14B	general
Qwen3 [94]	14B	general
DeepSeek-R1 [95]	14B	general
gpt-oss [96]	20B	general
Mistral Small 3 [97]	24B	general
Qwen3 [94]	32B	general
ALLaM [98]	7B	Arabic
Command R7B Arabic [99]	7B	Arabic
Jais [100]	7B	Arabic
SILMA [101]	9B	Arabic
BGE-M3 [102]	0.6B	embedding

zero-shot performance: the examples bias the model toward the demonstrated controls and away from the held-out target. That failure is consistent with an absence of leakage rather than with an experimental error. Under multi-label scoring, where the single-pick ceiling no longer binds, no model exceeds micro-F_1 0.32; models return too many labels rather than recover the curated mapping. Eleven of the twelve models completed the multi-label pass; the 32-billion-parameter model was terminated before completion.