



SHRIDHAR INFOSEC
S O L U T I O N S



AI-DRIVEN THREAT DETECTION · RESEARCH WHITEPAPER

Watching Both Surfaces

A Dual-Surface Detection Model for MSMEs without a SOC

SIS R&D Wing · Vadodara, Gujarat · Evidence-informed, not empirically validated

Watching Both Surfaces

A Dual-Surface Detection Model for MSMEs without a SOC

Research and Development Whitepaper

R&D Wing | Shridhar InfoSec Solutions (SIS)

Vadodara, Gujarat, India | Revision date: 30 September 2026

How claims are labelled. Evidence is a sourced observation or a stated legal or standards requirement. Interpretation is this paper's reasoned application of that material to a thinly staffed organisation. Proposal is an unvalidated design choice or implementation target.

This document is general cybersecurity guidance. It is not legal, regulatory, compliance, tax or process-safety advice. Organisations should confirm current obligations with qualified counsel and, where relevant, with CERT-In empaneled auditors.



Figure 1. Two in-scope surfaces on one desk: conventional IT, mail and identity; and AI use plus AI actions.

Contents

Page numbers are generated from the finished layout.

Abstract and keywords	4
Methodology	4
Executive summary	5
1. Introduction	6
2. Scope, MSME classification and Indian context	7
3. Two surfaces and a shared case workflow	8
4. Threats on both surfaces	10
5. Frameworks, directions and data-protection status	11
6. Operating constraint: shared IT, no internal SOC	12
7. Findings of this synthesis	12
8. Proposed Dual-Surface Detection Model	13
8.1 Watch	13
8.2 See AI use	14
8.3 Limit damage	15
8.4 Stay honest	16
8.5 Source-to-control matrix	16
9. Automation, approval and recovery	19
10. Baseline commitments and 90-day plan	21
11. Measurement and a proposed validation plan	23
12. Illustrative scenario	26
13. Proposed SIS delivery mapping	27
14. Limits, further work and close	27
Sources	28
Glossary and about SIS	29

Abstract

Organisations with shared IT responsibility and no dedicated internal 24x7 security operations centre now face two detection problems on the same desk. The first is familiar: endpoints, mail, identity and cloud. The second is newer: unofficial AI tools, model and data provenance, and agents that can write records, send mail or initiate payments. Major frameworks describe both problems. They do not, on their own, specify a first sequence that a thinly staffed Indian organisation can run without adding a second console nobody opens.

This paper is a targeted narrative synthesis, reviewed on 30 September 2026. It proposes a Dual-Surface Detection Model of four jobs - Watch, See AI use, Limit damage and Stay honest - and ten controls. The model is a translation of existing guidance into an operating profile, not a new detector and not a validated control set. It supplements, and does not replace, CERT-In's fifteen elemental cyber defence controls for MSMEs or the April 2022 directions on logging, time synchronisation and incident reporting.

Keywords. AI-driven detection; MSME; CERT-In; dual-surface model; prompt injection; agent permissions; DPDP Act; shared case workflow.

Methodology

This paper is a targeted narrative synthesis. It is not a systematic review. No independent second reviewer, no interview panel, no tenant trial and no statistical meta-analysis were performed.

Revision date. 30 September 2026. Claims in this version were checked against the documents listed in the source list. No systematic search protocol, second reviewer or interview panel is claimed.

Inclusion. Primary legal instruments and official guidelines applicable in India; widely used public standards and technical catalogues (NIST CSF 2.0 and SP 1300, NIST AI 100-2e2025, NCSC secure-AI guidance, OWASP GenAI LLM Top 10, ISO/IEC 42001 public overview, CISA phishing guidance); and two recent surveys that report Indian small-business samples with a stated method (CPA Australia Asia-Pacific Small Business Survey 2025/26; TTBS-CMR SME Digital Insights 2026: Cybersecurity). Vendor blogs and unsourced percentages were excluded.

Extraction. For each retained numerical claim the paper records publisher, year, sample or population, geography, what the percentage measures, and a direct document link. Where the original denominator or publication could not be verified, the claim was removed.

Synthesis. Standards and directions supplied principles. Survey evidence supplied operating context, not incidence rates for all Indian MSMEs. The four jobs and ten controls are the authors' proposed grouping of those principles for an organisation without an internal SOC. That grouping is labelled Proposal throughout.

Proposal. No live-tenant measurement of alert load, missed events or labour change is claimed.

Executive summary

The practical question is narrow. What should detection do in an organisation that has no security team, now that staff already paste work into public models and some tools can act?

Four findings follow from the synthesis. First, the organisation is defending two surfaces at once. Second, a tool that cannot land in the case workflow the firm already uses is likely to be ignored, whatever its accuracy slide says; that is not the same as saying every specialist console is a failure. Third, ten controls grouped into four jobs are a proposed starting baseline, not a proof that ten items are sufficient. Fourth, evidence-linked explanations for AI-use and AI-action events are proposed implementation requirements. This paper does not claim that no vendor currently provides them.

Indian obligations sit underneath the model. Covered entities must enable ICT logs continuously and retain them securely for a rolling 180 days; ninety days of searchable storage is not compliance on its own. Specified incidents must be reported to CERT-In within six hours of noticing them or being told of them. The Digital Personal Data Protection Act, 2023 is in staged commencement. A prompt that contains personal data is a data-handling risk that needs review. It is not, by itself, proof of an unlawful transfer or a reportable breach.

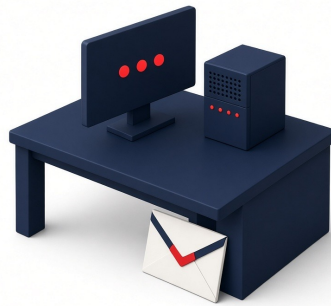


Figure 2. The intended operating profile is shared IT responsibility and no internal 24x7 SOC - not the legal MSME label alone.

1. Introduction

Shridhar InfoSec Solutions is a cybersecurity firm in Vadodara (shridharinfosec.com). Day-to-day work includes assessments, managed monitoring, ISO and SOC 2 support, DPDP readiness work, awareness testing and incident response. The Dedicated R&D Wing sits beside that delivery so research is not invented on a live ticket. One published focus on the R&D page is detection that a thin team can actually use.

A classic security operations centre does not shrink neatly. Many factories, clinics, regional financial shops and software firms will not staff analysts overnight. They buy a product, or a managed service, and they need that product to land on a desk that already runs payroll and vendors.

The same desk now has a second surface. People paste customer lists into public models. Developers pull components from the public internet. A small agent is given a key that can edit a record or raise a payment. NIST AI 100-2e2025 and the OWASP GenAI catalogue treat hidden instructions in retrieved documents, leaked secrets in replies, poisoned components and agents with excessive tool access as first-class problems.^{[8][10]}

The research question is therefore operational: what should AI-informed detection do in a firm with no dedicated security team, now that both surfaces sit on the same desk? Supporting questions: which controls reduce unproductive work without discarding useful effort; where existing instruments already speak; and what this paper is not allowed to claim in the absence of a measured tenant panel.

The contribution in Section 8 is a Dual-Surface Detection Model with ten controls, a baseline and a ninety-day planning horizon. It is secondary-research synthesis. It is not a new algorithm and it is not empirically validated.

2. Scope, MSME classification and Indian context

2.1 Official MSME classification

Under the Micro, Small and Medium Enterprises Development Act, 2006, as revised by Notification S.O. 1364(E) dated 21 March 2025 and applicable from 1 April 2025, enterprises are classified by a composite of investment in plant and machinery or equipment and annual turnover:^[15]

Class	Investment not more than	Annual turnover not more than
Micro	Rs 2.5 crore	Rs 10 crore
Small	Rs 25 crore	Rs 100 crore
Medium	Rs 125 crore	Rs 500 crore

Table 1. Official MSME thresholds from 1 April 2025. Source: Ministry of MSME / Budget 2025 revision.

That legal class is not the same thing as a staffing model. A medium enterprise at the top of the band can run a security function. A micro enterprise can still buy a managed SOC. This paper therefore separates the official label from the operating profile it is written for.

2.2 Intended operating profile

The intended reader is an organisation - often, but not only, an MSME - in which IT is a shared job, there is no dedicated internal 24x7 SOC, and the person who would open an alert also runs vendors, payroll or a helpdesk. The model may suit that profile. It requires adaptation for regulated sectors, multi-site plants, hospitals and any organisation that already runs a SOC or a formal ISMS.

2.3 What two recent surveys actually measured

CPA Australia's Asia-Pacific Small Business Survey 2025/26 was run in November-December 2025 across 4,166 businesses with fewer than 20 employees in eleven markets, including 513 Indian respondents. In that Indian sample, 47 per cent reported losing time or money to a cybersecurity incident in the prior twelve months, and 59 per cent said they had reviewed cybersecurity protections in the previous six months.^[1] That figure describes the surveyed Indian small businesses. It is not a census of all Indian MSMEs.

Tata Tele Business Services and CyberMedia Research published SME Digital Insights 2026: Cybersecurity (whitepaper dated in-file August-September 2026; public release 16 September 2026). The primary report is based on a survey of 800 IT decision-makers and business leaders across micro, small and medium enterprises in major Indian cities.^[2]

Two monitoring figures in that report measure different questions and must not be combined. On the frequency chart (How frequently do SMEs review cybersecurity?), 12 per cent selected continuous 24x7 monitoring, 18 per cent regular reviews with reporting, 35 per cent quarterly or half-yearly reviews, 26 per cent occasional or ad-hoc review, and 9 per cent review only after an incident. On the method chart (How cyber risk is monitored), 40 per cent selected periodic or manual checks, 28 per cent continuous monitoring and automated alerts, 16 per cent alerts and dashboards in some areas, 11 per cent no regular monitoring, and 5 per cent monitoring handled by an external IT partner.

Other primary-report figures retained here: 84 per cent planned to increase cybersecurity investment over the next 12-24 months; 40 per cent reported a cyber incident in the preceding 24 months; 45 per cent cited lack of expertise as the largest implementation barrier. The two surveys are not combined into one incident rate. Sample construction beyond the 800-respondent frame was not independently re-audited.

Evidence. Survey figures are directional context. They do not validate the ten controls.

3. Two surfaces and a shared case workflow

Detection is treated here as one operating problem with two places of work. They share a person. They rarely share a log format or a buying cycle.

3.1 Surface one - IT, mail and identity

Surface one is the laptop, mailbox, cloud console and identity system. Public reporting and joint guidance still treat phishing, reused credentials and ransomware as the events that take a small firm offline.^[5] AI is useful on this surface when it sits beside signatures and established rules, and when alerts land in the workflow the on-duty person already opens.

3.2 Surface two - AI use and AI action

Surface two has two parts that should not be collapsed. Use is a person pasting a file into a chatbot or connecting an unofficial plugin. Action is software calling a tool: writing a record, sending mail, changing a permission or initiating a payment. Detection that only sees the chatbot tab misses the agent. Detection that only sees the agent misses the paste.

3.3 Shared case workflow

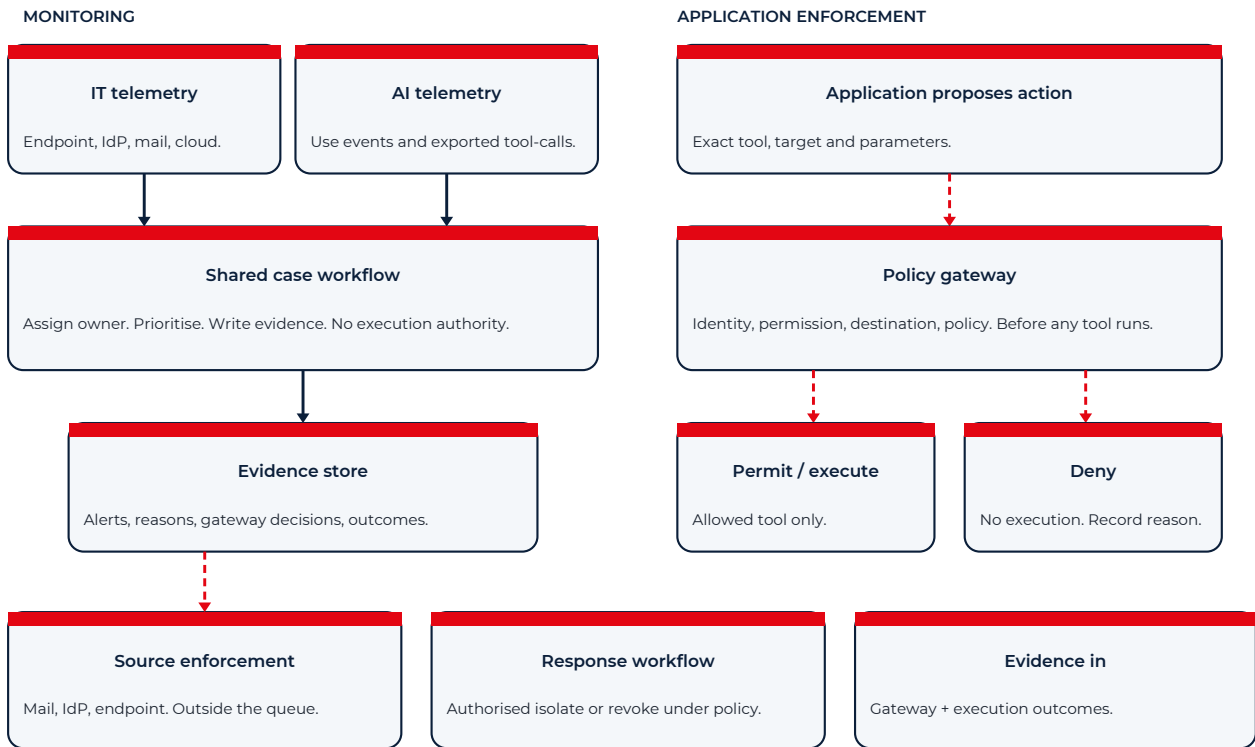
The design preference is one shared case workflow for both surfaces, with specialist consoles allowed where a control cannot be exercised from that workflow. A second console is a risk of ignored alerts. It is not automatically a failed programme.

Figure 3 is a reference architecture. It does not claim that every event is visible or that one product covers both surfaces.

Layer	What it collects	Where the person acts
IT telemetry	Endpoint, identity, mail-path and cloud events	Shared case queue
AI-use telemetry	Approved-tool inventory, browser or CASB hints, paste/upload events where available	Same queue, tagged AI-use
AI-action telemetry	Tool-call logs, retrieval sources, output destinations, approval records	Same queue, tagged AI-action
Enforcement outside the model	Vault, identity provider, mail gateway, application allow-lists	Preapproved playbooks plus human authorisation
Evidence store	Alerts with reasons, logs, approval records, monthly review notes	Retrievable for 180 days (see Section 5)

Table 2. Reference architecture. Visibility is incomplete wherever traffic is encrypted on unmanaged devices, personal accounts are used, or a provider will not export logs.

Solid navy = telemetry / evidence. Dashed red = proposed or authorised action. Queue does not grant authority.



Visibility gaps off these paths: unmanaged devices, personal accounts, providers that will not export logs.

Figure 3. Reference architecture. Monitoring lands in the shared workflow. Enforcement sits at source controls and at an application policy gateway. Visibility gaps remain.

4. Threats on both surfaces

The IT surface has not become uninteresting because models exist. A detection programme that starts with a novel AI use-case and leaves the mailbox alone has its priorities backwards.

Phishing, reused passwords and ransomware remain the events that take a small firm offline. Joint CISA, NSA, FBI and MS-ISAC guidance treats user instruction alone as insufficient; controls have to sit in the mail path.^[5] Voice and lookalike-domain lures are treated here as the same operating problem: someone pretending to be a person the staff already trust.



Figure 4. Phishing still arrives as mail, a lookalike domain, or a voice that sounds familiar.

Unofficial AI use does not wait for procurement. NCSC Shadow IT guidance treats unmanaged AI services used with corporate data as shadow IT. The related NCSC note on shadow AI describes staff turning to unapproved tools when approved systems do not meet the task.^{[9][17]} NIST AI 100-2e2025 catalogues related supply-chain and data-exposure issues.^[8]

Interpretation. A ban with no approved alternative often moves the same work onto unmanaged devices and personal accounts, which company telemetry usually cannot see. That is an operating inference, not a measured prevalence.

In this operating profile the following examples are illustrative: a founder drafts a tender in a public chatbot; accounts paste a debtor list to clean columns; a developer places an API key in a weekend experiment.

Prompt injection can influence a model's output. Whether anything executes depends on the application's tools and permissions. A summariser with no send tool cannot itself email invoices. Its output can still influence a person or another system that does have a send tool. Separating retrieved text from system instructions, and filtering inputs, are useful layers. They are not guaranteed prevention.^[10]

A compromised credential grants the scope that credential actually has - not automatically the entire service. An agent that can read untrusted mail and raise a payment is a payments path with a language model on the front. Read-only agents can still expose data.

Interpretation. OWASP category order is a catalogue structure. It is not an empirical ranking of attack frequency in Indian MSMEs.

5. Frameworks, directions and data-protection status

5.1 CERT-In directions, 28 April 2022

Direction No. 20(3)/2022-CERT-In, issued under section 70B(6) of the Information Technology Act, 2000, applies to service providers, intermediaries, data centres, body corporates and government organisations.^[3] Four obligations matter for this paper.

Logging. Covered entities must enable logs of all ICT systems and maintain them securely for a rolling period of 180 days within Indian jurisdiction, and produce them with an incident report or when directed. Logging is continuous. It does not start after suspicious activity. A design that keeps 90 days hot and searchable can be reasonable operations practice. It is not compliance unless the remaining required period is also securely retained, retrievable, and handled according to applicable jurisdiction and availability rules. This paper does not treat 90 days as sufficient.

Reporting. Specified Annexure I incidents must be reported to CERT-In within six hours of noticing them or being brought to notice. That clock is distinct from internal triage targets and from the monthly operational review in Control 10.

Point of contact. A designated person must be notified to CERT-In and kept current.

Clock synchronisation. ICT clocks must sync to NIC or NPL NTP, or to a source that does not deviate from them.

5.2 CERT-In MSME guideline

The official guideline index lists CISG-2025-03, 15 Elemental Cyber Defense Controls for Micro, Small, and Medium Enterprises (MSMEs), dated 1 September 2025.^[4] The document maps 45 baseline recommendations onto fifteen controls and is written against the official MSME classification. Some secondary write-ups have cited an older identifier; this paper follows the CERT-In index. Completing the ten controls in Section 8 does not establish CERT-In compliance. The ten controls supplement the fifteen-control baseline.

5.3 Digital Personal Data Protection Act, 2023

The Act received assent on 11 August 2023. Commencement notification G.S.R. 843(E) is dated 13 November 2025 and states that specified provisions come into force on the date of publication of that notification in the Official Gazette; further provisions come into force one year from that publication date; and the remaining listed provisions come into force eighteen months from that publication date. The e-gazette identifier associated with the published text is CG-DL-E-14112025-267647.^[4] This paper treats the operative clock as running from Gazette publication. It does not assert an independently verified calendar date for the eighteen-month point, because that date is an arithmetic consequence of the publication date rather than a second notified date. As of the 30 September 2026 revision, institutional provisions are in force; most substantive fiduciary obligations are not. This paper does not treat every DPDP duty as already applicable.

A prompt containing personal data creates a data-handling risk. It requires a review of purpose, provider terms, access, security, retention and any sectoral contract. It is not automatically an unlawful transfer or a reportable breach.

5.4 ISO/IEC 42001 and other catalogues

ISO/IEC 42001 is an AI management-system standard. This paper uses the public overview as context for inventory and accountability. It does not claim certification readiness or clause-level conformance.^[13] NIST CSF 2.0 and SP 1300 supply Detect and Respond language for small organisations.^{[6][7]} NCSC supplies secure-AI development principles.^[18]

6. Operating constraint: shared IT, no internal SOC

Labour is a design constraint, not a veto on every control that takes time. Some valuable controls add work: discovery, vaulting, compliant logging, restore tests and incident exercises. The test is whether the added work buys coverage, faster containment or a reduced blast radius that the firm can still operate.

A control that after thirty days still cannot be explained by the on-duty person, and that has no owner, is incomplete as an MSME control. That test is proposed. It is not a purchasing rule that any tool creating work is not worth buying.

7. Findings of this synthesis

Finding A - Evidence plus interpretation. The organisation now has two surfaces on one desk. Standards describe both. They do not sequence a first ten actions for a firm without analysts.

Finding B - Interpretation. Alerts that cannot be reached from the shared case workflow tend to be ignored. Prefer that workflow. Allow specialist consoles where the control cannot be exercised without them.

Finding C - Proposal. Ten controls in four jobs are a starting baseline. They are not a claim that ten items are enough for every sector or size.

Finding D - Proposal. Evidence-linked explanations for AI-use and AI-action events should be treated as implementation requirements. This paper does not assert that two explanations are universally missing from the market.

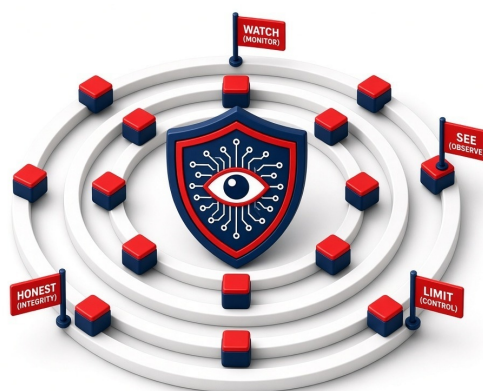


Figure 5. Four jobs around one operating problem. Ten controls sit on the jobs. The grouping is proposed.

8. Proposed Dual-Surface Detection Model

Each control is coded against a surface, a job and a labour effect. Labour effect is an expected operating burden, not a promise that the control is cheap. A control that after thirty days still has no assigned owner, no acceptance test and no retained evidence is incomplete as specified here. That completeness test is proposed.

Job	Controls	Surface	Labour effect
Watch	C1 shared workflow; C2 behaviour with signatures; C3 mail-path protection; C4 evidence-linked alerts	IT-Ops and IT-Human	Setup plus daily triage. Must not create a second unread inbox.
See AI use	C5 discovery and approved routes; C7 model and data provenance	AI-Use and AI-Act	Inventory is recurring work. Stale registers fail.
Limit damage	C6 scoped credentials; C8 input/output boundaries; C9 agent tool gates	AI-Act, with AI-Use on C8	Reduces blast radius when watching is late. Vaulting and reviews are not low-effort.
Stay honest	C10 logging, review and assurance	All	Compliant logging is continuous work. Monthly review is a calendar entry.

Table 3. Dual-Surface Detection Model. Funding Watch and skipping Limit leaves a live key after a late alert.

8.1 Watch

C1. Shared detection workflow and response ownership. Purpose: required in-scope sources reach a named responder. Owner: detection owner. Minimum: one queue, one roster, escalate-to path. Acceptance: from each required source, generate a harmless test event; confirm it arrives with timestamp, source evidence, severity and a next-action field; fail a connector on purpose and confirm that gap is visible. Opening ten old cases is not the test. Burden: medium. Isolate-host and disable-token only under Section 9. Disk wipe and payments stay out of software hands.

C2. Behaviour analysis alongside signatures and established rules. Purpose: detect attacks that evade signature and rule lists, without discarding those lists. Owner: detection owner with vendor or managed security service provider (MSSP). Minimum: signatures stay on. Acceptance: define selected use-cases, representative benign traffic and authorised anomalous tests; record detections, false alarms and missed tests over an initial two-to-four-week observation period. That period is observation, not automatic readiness to isolate. Guo reviews detection accuracy challenges when machine-learning methods move across changing environments; this paper uses that as a reason to keep signatures on, not as a measured result for Indian MSMEs.^[1]

C3. Mail-path protection and a monitored reporting route. Purpose: stop or contain phishing before the user is the only control. Owner: mail administrator. Minimum: gateway or inline controls, DMARC on owned domains, a report button a person reads. Acceptance: send a safe reported message and confirm the route to the responder the same day. DMARC authenticates the sending domain; it does not verify a voice call or a payment request. Joint CISA/NSA/FBI/MS-ISAC guidance treats user instruction alone as insufficient.^[5]

C4. Evidence-linked alert explanations. Purpose: the on-duty person can see observed facts, an evidence link, stated uncertainty and a justified next step. Owner: detection owner. Minimum: reason field plus link to the triggering record. Acceptance: ten sampled alerts have explanations that a reviewer can match to the underlying records; fluent wording that cannot be traced is a fail. Charmet et al. treat explainability as a condition for action, not a completed product feature.^[12] An AI model cannot independently lock accounts or revoke credentials. An authorised, preapproved response workflow may still isolate a host or invalidate a token under Section 9.



Figure 6. Watch prefers one shared queue. Specialist consoles are allowed when a control cannot be exercised from that queue.

8.2 See AI use

C5. Discover AI use and publish approved processing routes. Purpose: unofficial tools stop being invisible. Owner: named AI-use owner (may be the same person as detection). Telemetry: staff declaration, browser or Cloud Access Security Broker (CASB) signals where licensed, procurement list, mail-gateway file-type hints. Minimum: one page naming approved tools, banned paste classes, key rule, agent rule, owner and review date. Acceptance: a documented discovery pass was completed; available signals were reconciled with the inventory; every discovered exception was approved, blocked or recorded as deferred with an owner; visibility gaps were written down; a review date was set. A firm may have no unofficial tools. Finding an arbitrary number of problems is not the test. Burden: recurring, not low. Discovery will miss encrypted traffic, unmanaged devices, personal accounts and providers that do not export logs.

Field	Write it here
Approved tools	Named tools only. Unapproved tools do not receive company data.
Do not paste	Customer personal data, live deals, passwords, unpublished numbers, HR files, anything under an NDA - pending a documented exception.
Keys and identities	Workload identities or vaulted secrets. No personal API keys in chat.
Agents	No agent that can write, pay or send externally without recorded approval from [role].
Owner and review date	[name] / [quarter]
If unsure	Do not paste. Ask the owner.

Table 4. Minimum unofficial-use note. Fill the brackets. This is not a policy binder.

C7. Model, component and data provenance. Purpose: know what is running and whether it may see customer data. Owner: AI-use owner. Minimum: name, source, version, owner, data classes allowed. Acceptance: every production model or hosted API on the approved list has a row. If nobody can name the model, treat it as unofficial use under C5. NIST AI 100-2e2025 and OWASP LLM03/04 describe supply-chain and data-poisoning exposure.^{[8][10]}



Figure 7. See AI use is an approved list, a banned paste class, and a secret that does not live in chat.

8.3 Limit damage

Detection will be late some of the time. Limit is what remains standing.

C6. Scoped workload credentials and tested revocation. Purpose: a leaked credential grants only the access it was issued, and that access can be revoked. Owner: identity or IT owner. Minimum: one identity or secret per workload, vaulted, logged, with an expiry. Prefer managed workload identities where supported. Acceptance: where safely possible, disable the credential and confirm subsequent access failure, including cached sessions, secondary tokens and service dependencies. A tabletop tests preparedness; it does not prove technical revocation. Burden: medium to high. Compromised credentials grant their actual scope, not automatically the whole service.

C8. Input, retrieval and output boundaries. Purpose: untrusted text cannot quietly become a send, a write or a data dump. Owner: application owner. Minimum: retrieved text held apart from system instructions; retrieval only from authorised stores; destination restrictions; policy checks enforced outside the model. Acceptance: authorised tests of unauthorised retrieval, sensitive-output disclosure, a disallowed destination, an unapproved write, and a matching benign case to watch false blocks. Filtering and separation are layers, not guarantees.

C9. Agent tool restrictions and approval gates. Purpose: agents receive only the tools the job needs. Owner: application owner plus the approving role. Minimum: tool allow-list; start read-only where possible; human approval for payments, privileged access changes, external sensitive-data sends and bulk deletes. Approval binds to the exact action and parameters. Acceptance: an unapproved write is denied and produces a case with the tool name; a changed-parameter retry after an old approval is also denied. Read-only agents can still expose data.

8.4 Stay honest

C10. Logging, operational review and continued assurance. Purpose: detectors drift; logs must exist before the incident; someone must look. Owner: detection owner plus the CERT-In point of contact where designated. Minimum: ICT logs enabled continuously; rolling 180-day secure retention in line with Direction (iv); a working retrieve path; a monthly half-hour covering alert volume, justified triage, unofficial tools and the first key to revoke. Minimise sensitive prompt logging. Restrict access to retained AI-use evidence. Acceptance: test logging configuration, ingestion health, access control, retention setting and retrieval of the oldest sample the store can produce; write down what a new system cannot yet demonstrate across 180 days. A 30-day retrieve does not prove 180-day retention. The monthly review supplements, and does not replace, ongoing health checks and six-hour reporting. Burden: logging and retention are not low-effort.



Figure 9. Stay honest is a retrieve test and a dated review, not a dashboard score from last quarter.

8.5 Source-to-control matrix

Table 5 maps a source principle to a proposed implementation. It is conceptual mapping, not a formal compliance crosswalk and not evidence that implementing the control satisfies the source.

Control	Source principle	Proposed implementation
C1	NIST CSF 2.0 Detect/Respond; SP 1300 small-business start	Shared queue plus named owner
C2	Guo reviews accuracy challenges for ML detectors across changing environments (used as design caution, not a measured MSME rate)	Signatures stay on; observe before isolate
C3	CISA/NSA/FBI/MS-ISAC phishing guidance	Controls in the mail path plus a read report button
C4	Charmet et al. XAI survey; CSF Respond	Reason field linked to the triggering record
C5	NCSC Shadow IT / shadow AI guidance; ISO 42001 inventory context; CISC-2025-03 asset/data themes	One-page approved routes, quarterly look
C6	NIST AI 100-2; OWASP sensitive-info / agency; CERT-In logging of admin events	Vaulted scoped identity, tested revoke
C7	NIST AI 100-2 supply chain; OWASP LLM03/04	One-page provenance register
C8	OWASP prompt injection catalogue; NCSC secure AI	Retrieval authorisation, output checks outside the model
C9	OWASP excessive agency; NIST tool-using systems	Allow-list plus bound human approval
C10	CERT-In Direction (iv) 180-day logs; CSF Identify/Govern	Continuous logs, retrieve test, monthly review

Table 5. Source-to-control matrix. Conceptual only.

What a firm can start with existing systems is not the same as what needs a new connector. Reduced-scope options do not satisfy the full control.

C	Start with / owner	Needs extra	Reduced-scope (not full control)	Completion evidence / residual gap
C1	Mail/IdP consoles; ticket list. Owner: detection	Shared queue + AI connector	Hand-tag AI events in the IT ticket list	Harmless source tests + connector-fail alert. Gap: unmanaged devices
C2	Existing signatures. Owner: detection + vendor/MSSP	Behaviour analytics licence	Keep signatures; delay isolate	Logged detections vs authorised tests. Gap: drift
C3	Tenant anti-phishing + report button. Owner: mail admin	URL rewrite / sandbox	Turn on bundled protection and read reports	Safe reported message reaches queue. Gap: voice/payment
C4	Queue reason field. Owner: detection	Evidence-link field in the case	Write the reason by hand	Sampled explanations match records. Gap: missing telemetry
C5	Declaration + procurement. Owner: AI-use	CASB / browser connector	Posted list + quarterly declaration	Reconciled register. Gap: personal accounts

Table 6. Implementation options C1-C5. Reduced-scope does not satisfy the full control. Costs are not priced.

C	Start with / owner	Needs extra	Reduced-scope (not full control)	Completion evidence / residual gap
C6	Vendor-console identity. Owner: identity/IT	Vault / workload identity	Revoke the exposed secret in the console	Disable + access-fail test. Gap: cached sessions
C7	Named production APIs. Owner: AI-use	SBOM / component inventory	One provenance page	Every approved API has a row. Gap: transitive components
C8	App output filter if present. Owner: application	Policy gateway + retrieval ACL	Disable send/write tools	Unauthorised retrieval/disclosure tests. Gap: tools without logs
C9	Vendor allow-list if exposed. Owner: application + approver	Approval workflow + tool-call log	Read-only agent	Permit / deny / changed-parameter tests. Gap: silent tools
C10	Tenant retention setting. Owner: detection + CERT-In PoC	Searchable 180-day store	Export setting; retrieve oldest sample held	Setting + retrieve + gap note. Configured days ≠ accumulated history

Table 6 continued. Implementation options C6-C10.

9. Automation, approval and recovery

Four classes of action must be kept separate. An AI explanation never grants execution authority. Permission checks sit in the application or identity system, not in the model.

Class	What it may do	What it may not do
AI summary or recommendation	Draft a reason, rank a queue, suggest a playbook	Execute a high-impact action because the text sounds convincing
Deterministic enforcement	Block a known-bad hash, enforce an allow-list already approved	Invent a new action at run time
Preapproved response workflow	Isolate a host or disable a token under documented policy, tested triggers, notification and a rollback	Run against excluded critical assets, or without a recorded trigger
Human authorisation	Payments, privileged access changes, external sensitive sends, bulk deletes	Be reused if target, amount, recipient or payload changed

Table 7. Action-enforcement and approval. Reversible actions can still stop a clinic, a plant or a payroll run. Those systems need explicit exclusions and recovery steps.

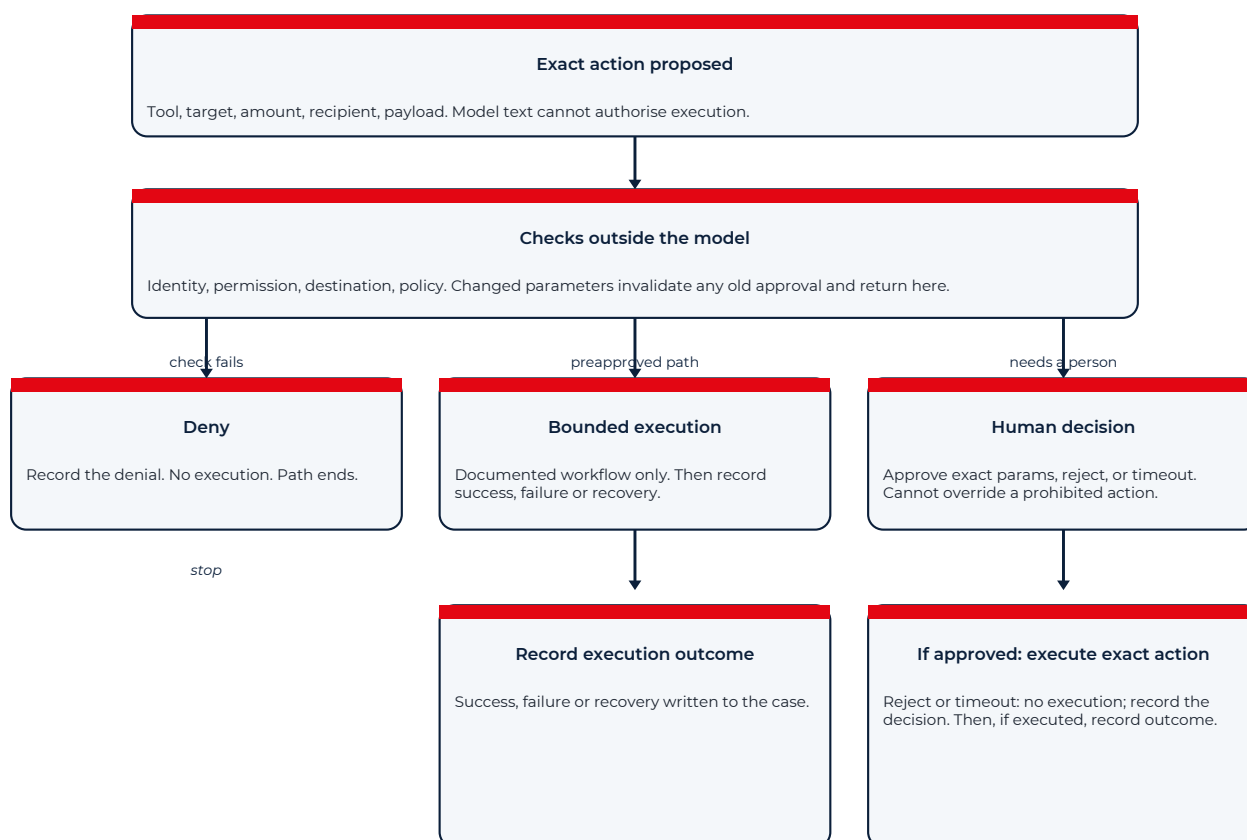


Figure 10. Action and approval flow. Failed checks deny. Preapproved work follows a bounded path. High-impact work needs a person. That person cannot override a prohibited action. Changed parameters restart the checks.

Hospitals, plants and other critical systems require a written exclusion list before any isolate-host workflow is enabled. An isolate that is technically reversible can still halt a production line or a ward system. Recovery steps - who lifts the isolate, from where, and after what check - are part of the control, not an afterthought.

Changed parameters void an old approval. A payment of Rs 12,000 to vendor A is not authority for Rs 12,000 to vendor B, or for Rs 1,20,000 to vendor A. The same rule applies to a bulk-delete count and to an external recipient list.

10. Baseline commitments and 90-day plan

The baseline does not replace MFA, patching, asset inventory, access control, backup restoration or incident preparation. Those remain foundational. The items below are the additional detection commitments this paper proposes. Ninety days is a planning horizon, not a guaranteed duration.

ID	Commitment	Job	Completion evidence
B1	Named owners for detection and for AI use; CERT-In point of contact if the entity is covered	See / Stay honest	Names on a page sent to the people who need them
B2	MFA on mail, cloud, VPN and AI consoles that support it	Watch / Limit	Console setting export or IdP report
B3	Mail-path phishing controls plus a monitored report button	Watch	Gateway policy and one handled report
B4	Shared case queue containing IT and tagged AI events	Watch	Harmless test events from every required IT and AI source arrive with timestamps, source evidence and required case fields; a simulated connector failure produces a visible health alert; ownership confirmed
B5	One-page AI-use note and first discovery pass	See	Posted note plus a dated list

Table 8. Baseline commitments B1-B5. Exposed credentials and dangerous agent permissions are handled immediately.

ID	Commitment	Job	Completion evidence
B6	Provenance row for every production model or hosted API	See	Register with owner and data class
B7	Workload identities or vaulted secrets; one revoke test	Limit	Revoke record and dependency check
B8	Agent tool allow-list and approval gate for write/pay/send	Limit	Permitted action succeeds; unauthorised action denied; changed-parameter retry denied; outcomes recorded
B9	ICT logs on; rolling 180-day retention design; retrieve test	Stay honest	Retention setting export plus retrieve of the oldest sample held, with gaps noted
B10	Restore test and a paper incident walk-through	Stay honest	Restore log and exercise notes

Table 8 continued. Baseline commitments B6-B10.

10.1 Days 1-15 - own, inventory, stabilise

Name owners. List laptops, cloud apps, AI tools and keys as they sit today. Turn on MFA where it already exists. Publish the one-page AI-use note. Prove yesterday's backup restores. If restore fails, remediate it while other essential work continues. Revoke any exposed personal API key on day one.

10.2 Days 16-45 - connect telemetry and protection

Stand up the shared queue. Put phishing controls in the mail path. Vault the practice-management or finance key. Connect endpoint and identity telemetry. Do not wait for a perfect CMDB.

10.3 Days 46-75 - test AI boundaries and tune

Complete a documented AI-use discovery pass and reconcile it with the inventory. Take write permission off any agent that does not need it. Start logging what a browser extension or retrieval pipeline sends, with prompt-minimisation. Tune behaviour detection. Run a safe prompt-injection simulation against a non-production connector that tests permission boundaries, sensitive-output disclosure and authorised retrieval, not only whether an email was sent.

10.4 Days 76-90 - exercise, evaluate, improve

Walk a stolen-token incident on paper, including the six-hour CERT-In decision if the scenario matches Annexure I. Hold the first monthly half-hour. Record what was missed. Decide what to stop.

Phase	Exit check
1-15	Owners named; restore result recorded; exposed keys revoked
16-45	Queue live; harmless source-event tests pass; connector-failure alert visible; mail-path control live; at least one key vaulted
46-75	Discovery pass complete; exceptions handled or deferred; permit, deny and changed-parameter tests recorded
76-90	Exercise notes; first monthly review; 180-day retention path described

Table 9. Phase exit checks. A missed exit check slips the horizon. It does not cancel urgent containment.

Items that are not deferred to a later phase: an exposed credential, an agent that can pay or send, a restore that already failed, and any Annexure I incident already in front of the team. Those are handled on the day they are found. Foundational MFA, patching and access-control work continue in parallel with the detection sequence. The ninety-day calendar does not replace them.

The calendar in Table 9 is a planning horizon, not a guaranteed deployment duration.

11. Measurement and a proposed validation plan

Fewer alerts alone do not prove improvement. Attackers can be quiet. A filter can hide events. Proposed measures:

Measure	Definition	Notes
Ingestion delay	Minutes from source alert creation to availability in the shared queue. Median and 90th percentile. Source clocks aligned to NIC/NPL practice where required. Missing source timestamps are reported as unknown, not zero	Data: source event id and queue ingest time
Triage delay	Minutes from queue availability to the first justified triage decision. Median and 90th percentile. Duplicates share one clock unless a new decision is required. Reopened cases start a new clock	Writing a sentence is not completed triage
End-to-end delay	Minutes from source alert creation to that decision, only where both timestamps exist	Do not mix clocks silently
Time to first explanation	Minutes from queue availability to the first evidence-linked explanation	Separate from triage delay
Actionable-alert rate	Reviewed alerts that warranted justified investigation or response / all reviewed alerts in the period	Not detection precision. Precision needs a ground-truth method the firm may not have
Explanation completeness	Reviewed alerts with the required evidence-linked explanation / all reviewed alerts	C4
Known-source coverage	In-scope systems sending usable logs / in-scope inventory	Stay honest
AI-register coverage	Observed tools with a register row / tools observed in the period	Not a count of all tools that may exist
Containment-test performance	Time and success of a documented revoke or isolate, including recovery	Limit C6
Approval-boundary outcomes	Denied, approved and exception results from write/pay/send tests, including changed-parameter retries	Limit C9
False blocks, missed tests, outages, disruption	User-reported false blocks; authorised tests not detected; connector health gaps; hours of interruption caused by a control	Honesty and harm
Responder effort	Internal staff hours on detection and AI-use review in the period	Setup, internal run-rate and provider hours logged separately

Table 10. Measurement scorecard. Targets are proposals. No performance gain is claimed. A quiet month is not improvement.

11.1 Proposed pilot protocol

This protocol has not been executed. No participating firm, incident, survey response or performance gain is reported here.

Questions. Does a shared queue receive harmless test events from each required source? Does a connector failure produce a visible health alert? Do allow-list and approval gates deny unapproved write, pay and send tests, including changed parameters, without an unacceptable false-block rate on declared benign tests?

Eligibility (proposal). Indian MSMEs under the investment-and-turnover definition, with mail and identity already in a cloud tenant, and with a named internal owner who can authorise safe tests. Exclude firms that cannot restore a backup or that already have an uncontained Annexure I incident.

Design (proposal). Four to eight consenting firms is a planning range, not a power calculation. A four-week baseline and an eight-week intervention give time to see connector health and a small set of authorised simulations; those lengths are planning assumptions. Firms may withdraw at any time. Company data stays in the firm's tenant.

Simulations. Predeclared, authorised and synthetic only: a reported test message; a harmless source event from each required connector; a connector-failure visibility test; an unapproved tool-call; a changed-parameter retry after an old approval; a benign allowed action to watch false blocks. Ground truth is the test plan plus independent reviewer adjudication.

Analysis and stop rules. Report Table 10 measures with denominators and missing-data counts. Confounders (staff change, season, new SaaS) are recorded, not adjusted away in this design. Stop a test that disrupts operations; recover under the firm's restore and revoke procedures. Results from volunteers cannot be generalised to all Indian MSMEs.

Protocol item	Proposed design choice
Primary outcomes (separate)	Ingestion pass-rate = required sources whose harmless test event arrives with timestamp, source id and case fields / required sources declared for that firm. Connector-visibility pass-rate = simulated connector failures that produce a visible health alert / simulated failures run. Do not hide one failure inside a combined score.
Secondary outcomes	Permit, deny and changed-parameter results; false-block count on declared benign tests; Table 10 delay and coverage measures
Repetitions (proposal)	Two authorised repetitions per required source and per action type. These are initial functionality checks, not evidence of reliability.
Mixed outcome	Both repetitions completed but results differ. Report both original results. Do not discard a fail after a later pass. A predeclared follow-up may retest after a recorded change; the original pair still stands. Mixed pairs are excluded from pass-rate numerators and counted in a mixed column.
Who agrees	The firm's named detection owner and the test lead sign a one-page pre-test form listing sources, tests, stop rules and any false-block cap before day 1.
Pass / fail / incomplete / disrupted	Pass: expected behaviour observed with required evidence. Fail: expected behaviour not observed. Incomplete: test not completed or evidence insufficient. Disrupted: test stopped because operations were affected. Mixed: completed repetitions differed.

Table 11. Proposed pilot protocol (part 1). Not executed.

Protocol item	Proposed design choice
Review	A second person not running the test scores the rubric. Disagreement is recorded; the stricter score is used unless both agree to reopen the evidence.
False blocks	False-block rate = declared benign tests blocked / declared benign tests run. Any cap is written on the pre-test form. It is a firm choice, not an industry threshold.
Missing data	Unknown timestamps stay unknown. Unavailable sources stay in the denominator of ingestion pass-rate and are marked unavailable.
Stop / recover / report	Stop a test that disrupts operations. Recover with the firm's revoke and restore steps. Report counts, denominators and missing-data notes. No significance tests in this design.
Data duties	Minimise payloads. Keep artefacts in the firm tenant. Delete synthetic artefacts at the end of the agreed period unless the firm asks to retain them.
Inference limit	Descriptive feasibility only. Not causal. Not representative of all Indian MSMEs.

Table 11 continued. Proposed pilot protocol (part 2).

12. Illustrative scenario

This section is illustrative. It is not a client case, not an SIS service feature list, and not evidence that the model has been validated.

A sixty-person accounting firm uses Microsoft 365, a practice-management platform and a public chatbot. In week one the AI-use owner finds a browser extension that pastes working papers into the chatbot, and an intern's personal API key on a shared drive. Those two items are handled immediately: the key is revoked, the extension is removed from company browsers, and the one-page note is posted.

In weeks three to six, endpoint and mail telemetry land in one queue. A safe lookalike-invoice test is quarantined, or its malicious URL is rewritten, in the mail path. The practice-management secret is moved to a vault and the previous token is invalidated. An agent tied to mail is set read-only.

A safe simulation drops a PDF into a non-production retrieval folder. The PDF contains an instruction to email recent invoices. The application policy gateway denies mail.send because that tool is not on the allow-list. The case reads: Denied tool mail.send from retrieval pipeline 'invoices-test'. Trigger: allow-list policy 'no-external-send'. Source document: sim-2026-09-01.pdf. Decision: deny. That sentence is only justifiable if the gateway emits tool name, destination, policy id and source document. The model's wording is not the control.

A second example case: AI-use upload of 'client-debtors.xlsx' to unapproved tool T. Detected by browser connector at 14:12 IST. User: A. Shah. Action: session blocked and owner notified. That sentence requires a connector that sees the destination and the file name. Many firms will not have that connector on day one. The scenario does not assume they do.

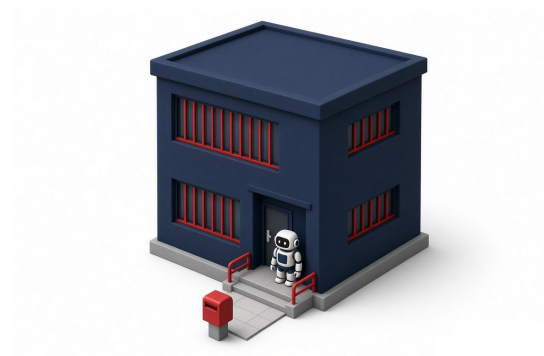


Figure 11. Illustrative firm only. Immediate work is the exposed key and the unofficial extension, not the ninety-day calendar.

13. Proposed SIS delivery mapping

The mapping below is a proposed way the model could sit on existing SIS service lines. It is not a verified catalogue, not an SLA, and not a claim that any named control is already implemented as a product feature.

SIS line	Possible contribution	Controls touched
Assessment	Surface inventory, unofficial-use discovery, provenance gaps	C5, C7, B1
Managed monitoring	Shared queue, mail-path events, evidence-linked reasons	C1-C4
vCISO / governance	Owners, AI-use note, monthly review, CERT-In PoC hygiene	C5, C10
Awareness	Paste discipline and report-button practice	C3, C5
Incident response	Revoke drills, six-hour decision tree, restore	C6, C10, B10

Table 12. Proposed delivery mapping only.

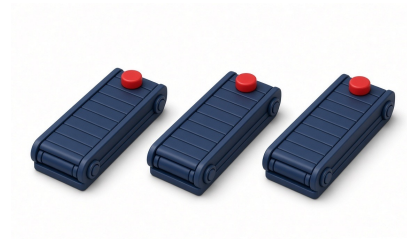


Figure 12. Four jobs can be staffed across assessment, monitoring and response. That staffing plan is proposed.

14. Limits, further work and close

This paper does not measure live tenants. It does not prove labour reduction. It does not make an organisation CERT-In compliant or DPDP-ready. Discovery is incomplete on unmanaged devices and personal accounts. Explanations are only as good as the telemetry underneath them. ISO/IEC 42001 is cited as management context, not as a completed certification path.

The next steps are practical validation and operational testing: a consenting-firm pilot using Table 10, a documented six-hour decision tree that a generalist can follow, and a tested revoke against cached sessions. None of that work has been completed for this revision.

The claim the paper is willing to defend is modest. Two surfaces now share a desk. Four jobs and ten controls are a coherent way to start. The rest is implementation, measurement and legal advice this document cannot give.

Sources

Numbered citations in the text point here. Homepages are not used as the location of a specific statistic.

[1] CPA Australia. Asia-Pacific Small Business Survey 2025/26. Fieldwork November-December 2025. 4,166 businesses with fewer than 20 employees in 11 markets; 513 Indian respondents. India finding: 47% lost time or money to a cybersecurity incident in the prior 12 months. Media release 28 May 2026 and full report. [cpaaustralia.com.au media release / survey page](https://cpaaustralia.com.au/media-release/survey-page)

[2] Tata Tele Business Services and CyberMedia Research. From Reactive to Real-Time: The Strategic Imperative for SME Cybersecurity / SME Digital Insights 2026: Cybersecurity. Primary whitepaper. Survey of 800 IT decision-makers and business leaders across micro, small and medium enterprises in major Indian cities. Frequency chart: 12% continuous 24x7 monitoring. Method chart: 28% continuous monitoring and automated alerts. Also 84% investment intent; 40% incident in 24 months; 45% expertise barrier. [tatatelebusiness.com whitepaper page](https://tatatelebusiness.com/whitepaper-page/) / PDF [primary PDF](#)

[3] CERT-In. Directions under section 70B(6) of the Information Technology Act, 2000, No. 20(3)/2022-CERT-In, 28 April 2022. Rolling 180-day ICT logs; six-hour reporting of Annexure I incidents; point of contact; NIC/NPL clock sync. cert-in.org.in/PDF/CERT-In_Directions_70B_28.04.2022.pdf

[4] CERT-In. 15 Elemental Cyber Defense Controls for Micro, Small, and Medium Enterprises (MSMEs), CISG-2025-03, 1 September 2025. [cert-in.org.in CISG-2025-03](https://cert-in.org.in/CISG-2025-03) and PDF [Elemental Cyber Defense Controls for MSME.pdf](#)

[5] CISA, NSA, FBI and MS-ISAC. Phishing Guidance: Stopping the Attack Cycle at Phase One. [cisa.gov phishing guidance](https://cisa.gov/phishing-guidance)

[6] NIST. Cybersecurity Framework 2.0. nist.gov/cyberframework

[7] Eliot, D. NIST SP 1300. The NIST Cybersecurity Framework 2.0 Small Business Quick-Start Guide. National Institute of Standards and Technology, 2024. [nvlpubs.nist.gov NIST.SP.1300.pdf](https://nvlpubs.nist.gov/NIST.SP.1300.pdf). Accessed 30 September 2026.

[8] National Institute of Standards and Technology. NIST AI 100-2e2025. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. 2025. [nvlpubs.nist.gov NIST.AI.100-2e2025.pdf](https://nvlpubs.nist.gov/NIST.AI.100-2e2025.pdf). Accessed 30 September 2026.

[9] National Cyber Security Centre. Shadow IT guidance. ncsc.gov.uk/guidance/shadow-it. Accessed 30 September 2026.

[10] OWASP GenAI Security Project. OWASP Top 10 for LLM Applications 2025 (Version 2025, dated 18 November 2024). Catalogue of issue classes (LLM01:2025-LLM10:2025), not an empirical frequency ranking. genai.owasp.org/llm-top-10. Accessed 30 September 2026.

[11] Guo, Y. A review of machine learning-based zero-day attack detection: Challenges and future directions. Computer Communications 198 (2023). DOI [10.1016/j.comcom.2022.11.001](https://doi.org/10.1016/j.comcom.2022.11.001). Also listed on [nist.gov publication page](https://nist.gov/publication-page). Accessed 30 September 2026.

[12] Charmet, F., Tanuwidjaja, H.C., Ayoubi, S. et al. Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications 77, 789-812 (2022). DOI [10.1007/s12243-022-00926-7](https://doi.org/10.1007/s12243-022-00926-7). Accessed 30 September 2026.

[13] ISO/IEC 42001. Artificial intelligence management system. Public overview only used here. iso.org/standard/81230.html

[14] Government of India. Digital Personal Data Protection Act, 2023 (Act 22 of 2023); Digital Personal Data Protection Rules, 2025; commencement notification G.S.R. 843(E), 13 November 2025. MeitY document set. [meity.gov.in DPDP Rules 2025](https://meity.gov.in/DPDP-Rules-2025)

[15] Ministry of Micro, Small and Medium Enterprises. Revised classification applicable from 1 April 2025, Notification S.O. 1364(E) dated 21 March 2025. msme.gov.in/know-about-msme

[16] Shridhar InfoSec Solutions. www.shridharinfosec.com and R&D shridharinfosec.com/rnd.

[17] National Cyber Security Centre. The hidden risks of shadow AI. ncsc.gov.uk/blogs/the-hidden-risks-of-shadow-ai. Accessed 30 September 2026.

[18] National Cyber Security Centre. Guidelines for secure AI system development. [ncsc.gov.uk secure AI guidelines](https://ncsc.gov.uk/secure-AI-guidelines). Accessed 30 September 2026.

Glossary and about SIS

MSME. An enterprise meeting the official investment and turnover thresholds in Table 1. Distinct from the operating profile in Section 2.2.

Operating profile. Shared IT responsibility and no dedicated internal 24x7 SOC.

Watch. The job of a shared case workflow, behaviour with signatures, mail-path protection and explained alerts.

See AI use. The job of discovering unofficial tools and keeping a provenance register.

Limit damage. The job of scoped credentials, input/output boundaries and agent approval gates.

Stay honest. Continuous logging, a retrieve test and a monthly review.

Write-call. An agent action that changes a record, sends mail or moves money.

CASB. Cloud Access Security Broker - a control point that can see or restrict use of cloud and browser tools.

MSSP. Managed Security Service Provider.

IdP. Identity Provider.

CMDB. Configuration Management Database - a system inventory. This paper does not require a perfect CMDB before detection starts.

RAG. Retrieval-augmented generation: an application technique that retrieves information and uses it to augment model generation.

Annexure I incident. A cyber incident type listed in the 28 April 2022 CERT-In directions, reportable within six hours of notice.

Evidence / Interpretation / Proposal. The three claim labels defined on the title page.

About the R&D Wing. Shridhar InfoSec Solutions publishes this note from its Dedicated R&D Wing in Vadodara. Sister paper: Closing the Translation Gap, an OT/ICS model for MSME industry. Neither paper is a substitute for a scoped engagement.

Shridhar InfoSec Solutions
Dedicated R&D Wing
Vadodara, Gujarat, India
www.shridharinfosec.com

Sister paper
Closing the Translation Gap
OT/ICS model for MSME industry
shridharinfosec.com/rnd



SHRIDHAR INFOSEC
S O L U T I O N S

DEFENDING YOUR DATA, SECURING YOUR SUCCESS

WATCH

SEE AI USE

LIMIT DAMAGE

STAY HONEST



Ready to take the next step?

Talk to the dual-surface detection team about a readiness conversation.

P

+91 93286 67642

@

info@shridharinfosec.com

W

www.shridharinfosec.com

SIS R&D Wing · Vadodara