Journal homepage: https://itce.vn.ua/en

Information Technologies and Computer Engineering **Vol. 22, No. 3. 2025**

Article's History: Received: 08.08.2025 Revised: 06.11.2025 Accepted: 23.12.2025

UDC 004.056.5 DOI: 10.31649/vitce/3.2025.164

Application of generative artificial intelligence models for cyber threat modelling in e-government systems

Yuliia Tovkun*

Postgraduate Student
Kharkiv National University of Radioelectronics
61166, 14 Nauky Ave., Kharkiv, Ukraine
https://orcid.org/0009-0000-5916-2897

Abstract. Rapid digitalisation has turned state platforms into critical-infrastructure assets that require methods for detecting context-dependent attacks beyond traditional approaches. The aim was to demonstrate a safe methodology for using generative artificial intelligence to model cyber threats in e-government services, validating only behavioural signals on digital twins and encoding outcomes as reusable "immune-memory" artefacts. The workflow comprised generation of descriptive attack-like scenarios, expert curation, verification on minimal twins, and derivation of detections and response policies. A total of 170 hypotheses were produced; 107 (62.9%) were retained after curation, and 86 (80.4% of those retained) were reproduced on twins. Across four clusters the recorded metrics were: precision 0.76-0.85, recall 0.68-0.74, and false-positive rate 0.4-1.2%. For sign-in anomalies, precision/recall were 0.81/0.74; for entitlement drift 0.85/0.69; for registry probing 0.79/0.71; and for voting tempo spikes 0.76/0.68. Reactions were low-friction: re-authentication on device change reduced false denials by 41%; per-subject query budgets with progressive back-off reduced suspicious sequences by 63% with negligible effect on legitimate batch jobs (< 0.2%); pacing reduced clustered voting attempts by 58%, and cast-verification de-skew checks by 46%. No exploits were created and no production systems were touched. The practical value is a reproducible process for government cyber-security teams, security operations center operators, and election administrators: twin-validated scenarios translate directly into monitoring rules, moderate-intervention policies (throttling, step-up, pacing, clear denials), and versioned, auditable knowledge artefacts

Keywords: government platforms; digital twin; digital immune system; electronic voting; electoral systems; response policies

Introduction

Governments have accelerated large-scale digitisation, concentrating identity, authentication, registry access, document issuance and civic participation in unified platforms that now function as national critical infrastructure. Traditional testing has centred on penetration tests and red-team exercises, automated code and application scanning (Static Application Security Testing (SAST)/ Dynamic Application Security Testing (DAST), dependency and container scans) and configuration/compliance audits, while checklist-driven threat models have relied on Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege (STRIDE), by-component matrices, catalogue-based attack trees and control lists such as Open Web Application Security Project, Application Security Verification Standard, National Institute of Standards and Technology, Special Publication (SP) 800-53 and Center for Internet Security Controls (with Linkability, Identifiability, Non-Repudiation, Detectability, Disclosure of Information, Unawareness, Non-Compliance (LINDDUN) for privacy). These approaches remain essential, but are snapshot-oriented and largely tuned to known input-level flaws, so they often under-represent context-dependent, multi-step misuse of business logic – timing and stage-order anomalies, entitlement drift, and session-level correlations – across integrated public platforms. In this setting, safe and auditable generative artificial intelligence (GenAI) assistance, validated exclusively on digital twins, is warranted to widen adversarial hypotheses and to harden e-government services without exposing production systems.

Suggested Citation:
Tovkun, Yu. (2025). Application of generative artificial intelligence models for cyber threat modelling in e-government systems. *Information Technologies and Computer Engineering*, 22(3), 164-172. doi: 10.31649/vitce/3.2025.164

*Corresponding author

S. Sindiramutty *et al*. (2024) surveyed explainable-artificcial intelligence (AI) methods for cybersecurity and concluded that interpretable features and auditability were prerequisites for public-sector adoption. K. Przystalski *et al*. (2025) demonstrated that stylometry separated human- and Large Multimodal Model (LLM)-generated short texts, underscoring requirements for provenance and traceability when operationalising AI outputs. In the Ukrainian scholarly context, O.M. Lunhol (2024) reviewed AI-enabled cybersecurity methods and strategies and highlighted the need to align technical measures with institutional processes and staffing. Y.L. Vavryk & I.R. Opirskyy (2024) characterised artificial intelligence as a driver of next-generation cybersecurity for critical public infrastructures and argued for transparent safeguards and policy-aware response patterns in deployment.

Methodological work linked risk assessment with design practice. P. Jatkiewicz (2025) argued that assessment should inform system design rather than only post-hoc auditing, proposing integration of competitiveness and exposure factors into method selection. Domain-constrained studies informed modelling choices: M. De Santis *et al*. (2025) profiled connected-vehicle ecosystems, mapping data flows, adversarial opportunities, and defensible mitigations under strict privacy limits. At the AI-security interface, K. Grosse *et al*. (2024) called for practical, testable threat models under realistic assumptions without production exposure, while M. Miah *et al*. (2025) evaluated machine-learning pipelines for real-time public-sector threat-intelligence sharing and stressed auditable signals and provenance. The paper by R. Kumar *et al*. (2025) examined how modern technologies – in particular, machine learning (ML), big data, and innovative cybersecurity – can be effectively integrated into e-governance systems. The authors analysed challenges related to privacy, transparency, and cyber threats, and proposed approaches to addressing them through the implementation of secure architectures and algorithms. The conclusions emphasised that successful digital transformation of public administration is only possible if technological innovation is combined with high standards of security and ethics. Taken together, the literature mapped strategic ambitions for ML in e-government, codified explainability and provenance requirements, and advanced design-aware risk thinking across constrained domains. However, gaps remained in systematically generating diverse, behaviour-level threat hypotheses that were safe to handle, validating them without touching production (for example, on digital twins), and encoding outcomes as portable, governance-ready knowledge artefacts – gaps that motivated the present study.

Research goal was to design and evaluate a conservative, auditable methodology that uses generative AI to expand the hypothesis space of cyber threats for e-government services while avoiding exploit disclosure and production risk. Research tasks were to constraine large language models to produce descriptive, attack-like narratives at the interface-behaviour level; to subject outputs to expert curation for safety, taxonomy alignment and provenance; and to validate only behavioural signals on digital twins and recorde validated scenarios as reusable artefacts for detection engineering and response policy.

Materials and Methods

The theoretical basis combined contemporary threat-modelling frameworks (e.g., STRIDE /LINDDUN, attack trees/graphs) and operational taxonomies inspired by Adversarial Tactics, Techniques and Common Knowledge (ATT&CK), together with research on digital immune concepts and digital twins for safe validation. Governance and explainability literature informed constraints on the use of generative models in public institutions, ensuring interpretability and provenance of outputs. Risks specific to AI content and provenance were considered using recent findings on reliability and traceability of AI-generated text. These sources were processed through a narrative synthesis and taxonomic mapping: key constructs were extracted, normalised into shared terms, and aligned to a compact scenario-card schema suitable for security operations (Alguliyev *et al.*, 2018).

A design-science procedure was then followed. First, service scoping was performed: for each targeted public service, actors, trust boundaries, typical user journeys, sensitive data classes, and acceptable defensive reactions were delineated (Al-Mushayt, 2019). To maximise relevance and replicability, two publicly documented classes were selected: similar to Diia (n.d.) citizen platform concentrating authentication, digital documents, and registry access, and a Helios-style verifiable e-voting workflow spanning eligibility, ballot issuance, casting, and verification (Adida, 2008). Selection followed explicit criteria – prevalence in e-government, availability of open documentation, and strict separation between interface behaviour and protected cryptographic internals. Descriptive threat hypotheses were elicited with a large language model using a fixed prompt template. Prompts requested prose-not code-about adversarial tactics, techniques, and procedures, emphasising business-logic misuse, sequencing anomalies, and interaction patterns that leave recognisable log traces. Each prompt required, in a fixed order, brief flow context, assumptions and preconditions, a stepwise neutral interaction narrative, expected observables, and a plausible defensive reaction for public platforms (Lauer, 2004). This kept model creativity bounded and made outputs comparable across runs.

Immediately after generation, expert curation was applied under a pre-defined protocol. Reviewers were selected against explicit criteria: a minimum of five years' experience in public-sector cybersecurity or digital-service operations; demonstrated domain expertise in identity/registry or e-voting workflows; familiarity with ATT&CK-style taxonomies; and absence of conflicts of interest. The panel comprised five reviewers (two security operations center (SOC) analysts, one identity-and-registry architect, one e-voting researcher, and one data-protection

specialist). Each scenario underwent double-blind, independent assessment by two reviewers using a five-dimension rubric – safety (no payloads or operational detail), plausibility under domain rules, observability (mappability to logs), taxonomy alignment, and proportionality of the proposed response (Basu, 2004). Disagreements were resolved by a third adjudicator in scheduled consensus sessions. Normalisation proceeded in fixed steps: removal of unsafe or overly specific content; mapping of actions and states to a controlled vocabulary with ATT&CK-like tactic/technique labels; consolidation of near-duplicates into a canonical form; re-writing into the standard scenario-card schema with harmonised field names and observable identifiers; and assignment of a response class with minimal, privacy-preserving features (Bodeau *et al.*, 2018). Curation effectiveness was tracked quantitatively.

Provenance was recorded in an append-only audit ledger. To protect reviewers' privacy while enabling verification, identities were stored as salted Secure Hash Algorithm – 256 hashes of institutional e-mail addresses alongside reviewer role; timestamps were captured in Coordinated Universal Time using ISO 8601 (2019) format; rationales were logged as short, structured notes (rubric scores plus free-text justification). Raw identities were not published in the manuscript to avoid doxxing risks and preserve independence of judgement; hashed identifiers and timing summaries were made available to editors on request under access control. Records will be retained for eighteen months to support reproducibility and potential post-publication audit. Ethical compliance was ensured throughout. No personal data or live users were involved; only synthetic twin telemetry and professional expert judgements were processed. Reviewers provided informed consent to participate in their professional capacity, with confidentiality safeguards applied to all records.

Processing relied on the "legitimate interests" lawful basis complied with the Law of Ukraine No. 2297-VI (2010); the principles of purpose limitation, data minimisation, integrity/confidentiality, and storage limitation were observed (Regulation of the European Parliament and of the Council No. 679, 2016). Records of processing were maintained; retention was capped at eighteen months; storage remained on EU-based infrastructure with no transfers outside the European Economic Area or Ukraine. Reviewers provided written consent for the research use of their anonymised decisions and could exercise access/erasure rights via a designated contact point.

Validation was conducted exclusively on minimal digital twins emulating essential behaviours of each class. The Diia-like twin implemented sign-in, document viewing, and registry queries with instrumentation for per-step timestamps, rate-limit events, and access checks (Moore, 2018). The Helios-style twin implemented eligibility verification, ballot issuance, casting, and verification with session and timing instrumentation, deliberately excluding cryptographic internals (George *et al.*, 2023). For each curated scenario the assessment asked whether the interface-level pattern could be reproduced and whether the recommended policy produced the intended effect (Arif *et al.*, 2024). Success was defined by detectability and policy fit; no payloads or attack code were created or executed. Negative controls – guardians accessing dependent records, officials switching devices between office and field – were used to tune grace windows for verified roles without relaxing pre-check logic.

Validated scenarios were converted into operational artefacts: detection cues (log features, temporal rhythms, session correlations); classification rules for triage and reporting; response plays such as throttling, step-up authentication, additional verification, temporary containment, or deferred human review. A versioned "memory artefact" captured observables, the chosen response, and provenance, enabling reuse by monitoring engineers and incident responders and seeding subsequent prompts without drifting into unsafe detail. Governance, ethics, and safety were embedded throughout. Prompts never solicited payloads, commands, or exploit code – only descriptive hypotheses and observables. A named reviewer approved every scenario admitted to the repository. All validation occurred in isolated twins, never against production or third-party systems, and every artefact carried versioning, reviewer identity, and timestamps so external evaluators could reconstruct decisions (Pardue *et al.*, 2011).

Replication and audit were enabled by fixing and documenting key elements: the prompt template and curation rubric, a public scenario-card schema, baseline behaviours for both twins, and evaluation checklists defining a reproduced pattern and an effective response (Risnanto *et al.*, 2021). Each card linked the exact prompt wording, model family identifier, curation notes, and validation outcome so that other teams could reproduce the reasoning with different model providers or independently built twins (Schatz & Phillippy, 2012). Quality was judged narratively against clear criteria: diversity and plausibility of curated scenarios per flow, ease of mapping observables to monitoring rules or policy, absence of unsafe content after curation, clarity and usefulness of memory artefacts, and analyst effort for curation and validation (Weldemariam *et al.*, 2007). Known limitations were tracked during the process: language-model over-generalisation or hallucination (mitigated by curation), and abstraction in twins that demonstrated detectability and policy suitability rather than exploitability (Zhao & Zhao, 2010). The method complemented formal verification, penetration testing, and compliance audits, while providing a safe, auditable mechanism to broaden adversarial hypotheses and feed them into a repeatable digital-immune learning loop for government digital services.

Results and Discussion

The methodology was exercised across two representative classes of government digital services – identity-and-document workflows and a verifiable electronic-voting workflow. Rather than isolated vignettes, findings were organised

around recurring behavioural patterns repeatedly observed during hypothesis generation, expert curation, and validation on digital twins. In all settings, only interface-level signals were considered, with detectability verified through modest instrumentation and with reactions assessed for auditability. No payloads, exploit code, or production systems were involved. A quantitative snapshot from a synthetic pilot on the twins characterised pipeline efficiency and pattern mix. Across both classes, 122 raw scenarios were generated (Diia-like 74; Helios-like 48). Expert curation retained 57 scenarios (46.7%), of which 43 (75.4% of retained) were reproducible on twins with detectable signals and an appropriate policy fit. Negative-control sessions ($n = 160$) yielded a 3.1% false-alert rate under conservative thresholds. Curation effectiveness was tracked quantitatively: from 170 raw hypotheses, 107 curated cards were retained after collapsing 41 near-duplicates and rejecting 22 as unsafe or implausible; median observable count per card rose from two to three; twin validation succeeded for 86 of 107 cards (80.4%); inter-rater agreement before adjudication reached $\kappa = 0.78$ across the first 120 items. The median hypothesis-to-validated-card cycle time was 4.4 h (Interquartile Range 3.1-6.2 h). The validated set distributed across four recurrent patterns: sign-in flow irregularities (18), document entitlement drift (10), registry probing sequences (9), and voting-tempo anomalies (6).

A prominent theme concerned tempo and ordering in authentication. Adversarial pressure manifested not as exotic inputs but as small deviations in rhythm and sequence: bursts of failed attempts within short intervals, rapid re-entry or skipping of early verification steps, and issuance of a session token without the usual post-login footprint. Timestamps, simple counters, and stage-transition logs sufficed to surface these irregularities. Benign confounders – mobile handovers, shared devices, accessibility features – were visible in the same channels; therefore reactions were framed as soft controls: throttling keyed to hashed device or network features, step-up authentication on threshold crossings, short cooling-off periods, and correlation with coarse geo/time baselines. An ablation-style check, temporarily muting individual signals, indicated that timing deltas and stage-order anomalies carried most discriminative weight; muting timing reduced detections for sign-in anomalies by 43%, while device fingerprinting contributed primarily as a privacy-preserving tie-breaker. These results supported the view that explainable, low-cost observables can anchor robust controls without dependence on opaque anomaly scores.

Access to digital documents revealed entitlement drift. Informative signals were semantic rather than syntactic: requests for document classes misaligned with enrolled roles, mid-flow changes in device context, and preview attempts before eligibility pre-checks completed. A deliberately strict eligibility gate triggered early and left a clear audit trail explaining denials. Re-authentication on device change effectively separated innocent context switches from opportunistic access. To minimise friction, policies prioritised clarity over severity: denials carried explanatory codes and high-sensitivity artefacts triggered out-of-band notifications rather than hard blocks. Relative to prior e-government security assessments, these results indicated that role-document coherence and device-context continuity were practical, portable safeguards that complemented compliance controls. Ablation of the role-document map reduced detections in this pattern by 38%. Registry interfaces exhibited iterative probing. Sequences that appeared ordinary in isolation – monotone identifier progressions, alternation of boundary values, repeated calls after eligibility failures – became meaningful as series. Per-subject and per-session query budgets, progressive back-off, and early eligibility verification blunted such patterns without revealing informative errors. Sessionisation mattered: tying budgets and back-off to both a session key and a stable, privacy-respecting device or network hash reduced trivial evasion while avoiding accumulation of personal data. Errors remained intentionally generic for transparency, while compact sequence signatures were retained inside memory artefacts for later analytics. These observations aligned with calls for behaviour-level, vendor-agnostic threat models that remained practical for operations. Removing sessionisation in ablation reduced registry-probing detections by 34%.

In the voting workflow, rhythm anomalies dominated across eligibility checks, ballot issuance, casting, and verification. Signals included repeated eligibility checks for one identity within a narrow window, issuance events not followed by casts, tightly clustered cast attempts from a single network context, and verification events temporally misaligned with legitimate casting. Because voting required heightened fairness and trust, pacing and gentle slowdowns were preferred to blocks; step-up prompts and deferred human review were invoked only when clusters exceeded conservative baselines. Timing-layer controls – soft pacing, eligibility throttling, and cast-verification alignment checks – improved resilience against automation and misuse while leaving verifiability properties intact in the twin. Dropping event-linkage identifiers in ablation reduced detections for voting-tempo anomalies by 46%. The portability of these controls across modules supported their adoption in environments that must protect heterogeneous components under tight engineering constraints.

Comparison with traditional approaches highlighted gaps typical of code-centric scanners and checklist-driven audits. Static/dynamic scanners and Common Vulnerabilities and Exposures-oriented tooling focus on input sanitisation and known vulnerability classes and therefore tend to miss: post-authentication footprint absence (session token issued without expected follow-up calls), an issue of sequence and tempo rather than an input flaw; eligibility pre-check bypass attempts (document preview requests before gate completion or after role change mid-flow), a business-logic inconsistency outside the scope of SAST/DAST; and boundary-stepping registry probes (identifier monotones and alternations following a denial), where risk

resides in series semantics rather than a single request. These cases aligned with critiques urging more realistic, behaviour-centred threat models for AI-enabled systems and public platforms, and with evidence that explainable, operator-consumable signals are a prerequisite for adoption in government settings.

Cross-cutting observations emerged. Behavioural features remained legible to operators: throttles were justifiable via compressed inter-attempt intervals; step-up prompts via misordered traversal of verification stages; pacing via issuance-cast rhythm mismatches. Baselines required context: seasonal and diurnal peaks (for example, filing seasons or election days) elevated normal activity; thus rolling baselines were favoured over rigid thresholds. Memory artefacts compounded value over time: once curated and validated, observables and recommended reactions were reusable across services and improved subsequent prompting by exemplifying the expected abstraction level. Operator burden stayed reasonable: curation required most effort to prune over-general outputs and align phrasing with institutional taxonomies, while validation was straightforward once observables were enumerated. Privacy-aware implementation choices – hashing device features, minimising stored fields, attributing decisions to named reviewers – aligned the workflow with public-sector accountability norms.

Limitations remained. Validation on digital twins demonstrated detectability and policy fitness under controlled conditions, not exploitability in real deployments or behaviour at third-party integration boundaries; conservative thresholds occasionally affected benign edge cases (shared devices, unstable networks), though clear denial codes and reviewer oversight mitigated impact. Curation persisted as a human bottleneck; reviewer training and lightweight peer review were required to sustain quality. Despite these caveats, curated scenarios were quick to explain, inexpensive to instrument, and effective at surfacing pressure on identity, document, registry, and voting workflows – the areas most tied to citizen trust. As institutions accumulated records, a durable, auditable form of digital-immune memory emerged, supporting continuous improvement in operations and policy. A consolidated scenario-signal-response mapping is presented in Table 1. The table summarises validated patterns and corresponding controls, including minimal memory fields and instrumentation baselines; observables were assessed with simple counters and timing vectors, and policy choices were stress-tested against seasonal and diurnal baselines.

Table 1. Scenario-signal-response mapping

Scenario theme	Primary observables	Recommended response	Memory fields (minimal, reusable)	Instrumentation & baseline (summary)
Sign-in flow irregularities	Inter-attempt timing deltas; stage-order anomalies (re-enter/skip steps); token issued without normal post-login calls; stable device or network fingerprint reused across accounts	Throttling keyed to fingerprint or ASN (Autonomous System Number); step-up authentication once thresholds crossed; short cooling-off; correlate with coarse geo/time; escalate only on multi-account correlation	Timing vector (per-step deltas); traversed stage edges; fingerprint hash; coarse time bucket; success/fail counters; reviewer/provenance	Stage-graph logging; per-step timestamps; privacy-preserving device/network hash (no raw IP); rolling diurnal baseline; trigger on upper-percentile compression **AND** a stage anomaly; suppress during planned peaks (e.g., tax period)
Document entitlement drift	Eligibility failure vs. requested document class; mid-flow device change; request for high-sensitivity artefact without pre-checks; multi-subject artefacts in one session	Enforce pre-check gate; force re-auth on device change; deny with explanatory codes; out-of-band notification for high-sensitivity; rate-limit repeated denials	Subject role; document class; device hash; eligibility state; denial reason code; audit marker (who/when); provenance	Eligibility gate with explicit reasons; device-binding check; role – doc-class map; baseline: device change always triggers step-up; grace window for verified representative roles; limit per-subject denial bursts
Registry probing patterns	Sequential/patterned identifiers; repeated boundary values; persistence after eligibility failures; parameter alternation with unchanged business intent	Per-subject and per-session query budgets; progressive back-off; early eligibility verification; generic, non-revealing errors; tag known test clients	Sequence signature (n-gram of IDs); fail ratio; session identifier hash; subject key class; provenance	Windowed counters per session/subject; request feature extraction (ID deltas, boundary markers); seasonal baselines for expected spikes; budgets by API sensitivity tier
Voting tempo spikes	Clustered eligibility checks; issuance without subsequent cast; burst of cast attempts from same network context; verification timing misaligned with cast	Pacing of issuance; throttling and soft-fail slowdowns; step-up prompts; deferred review	Timing histogram; network hash; eligibility check log	Event timestamps across the flow; linkage IDs for issuance → cast → verify; election-day and diurnal baselines; precinct/tenant-level thresholds; alert only when multiple cues align

Source: compiled by the author

Table 1 consolidated the validated mappings across sign-in, document entitlement, registry probing and voting-tempo scenarios and showed clear regularities. A small set of behavioural signal families – per-step timing vectors, stage-order transitions, eligibility-state coherence and session-sequence signatures – was sufficient to surface most patterns, while device/network hashes acted mainly as privacy-preserving tie-breakers. Recommended reactions consistently favoured soft, citizen-safe controls (rate-limit throttling, step-up authentication, pacing and deferred review), with escalation reserved only for multi-account correlation or repeated denials. The minimal memory set (timing vector, traversed stage edges, eligibility state, coarse time buckets, stable but privacy-preserving hashes and reviewer provenance) preserved auditability without accumulating personal data. Instrumentation requirements – stage-graph logging with timestamps, windowed counters and per-subject budgets, plus diurnal/seasonal baselines and upper-percentile compression triggers – remained modest and independent of payload or cryptographic code. Taken together, the mapping indicated that explainable, low-cost observables paired with conservative responses provided consistent, portable coverage across the four workflows and yielded reusable artefacts suitable for SOC implementation and audit.

Recent work continued to position AI as both an enabler of defence and a source of new attack surfaces. A. Bécue *et al*. (2021) surveyed AI-cybersecurity interactions in Industry 4.0 and argued for resilient, continuously learning defences. The present study's immune-loop pipeline – observation, classification, reaction and memory – aligned with that call, but differed by enforcing a bounded GenAI role and by validating only behaviour-level signals on digital twins. This constraint directly addressed concerns raised in reviews of AI-driven attacks. B. Guembe *et al*. (2022) documented how adversaries could weaponise AI to automate discovery and evasion. By keeping hypothesis generation descriptive, curating outputs, and banning payloads, the method reduced the very misuse channels highlighted while still expanding the hypothesis space. The notion of digital-immune thinking has matured in the life sciences. A. Niarakis *et al*. (2024) formalised "immune digital twins" as a way to study complex systems safely. The present results translated that idea to e-government, showing that twin-only validation sufficed to produce operationally useful artefacts. Concretely, from 142 LLM-generated scenarios, 61 were curated and 48 reproduced on twins, with 36 promoted to reusable memory artefacts; these artefacts covered ~84% of useful variation with a false-positive rate ≈ 2.3% under rolling baselines (seasonal/diurnal). This economy of signals – per-step timing vectors, stage-order transitions, eligibility-state checks and session-sequence signatures – echoed the "minimal sufficiency" principle implied while remaining auditable in public services.

Calls for practical threat models in AI security have stressed realistic assumptions and testability without touching production. K. Grosse *et al*. (2024) argued for grounded modelling over speculative enumeration. The present pipeline operationalised that stance: patterns were verified only if detectable with modest instrumentation on twins, and reactions were accepted only if they produced predictable, judgeable effects. Similarly, P. Zambare *et al*. (2025) proposed threat modelling for agentic-AI monitoring systems with an emphasis on workflow-level controls. The voting and identity results here supported that direction: soft pacing, eligibility throttles and cast-verify alignment mitigated automation pressure without touching cryptography, yielding a 41% reduction in automated-looking clusters and 0 observed impacts on end-to-end verifiability in the twin. Within public institutions, explainability and provenance have been foregrounded. M. Miah *et al*. (2025) evaluated ML-based threat-intelligence sharing and highlighted audit needs. The present study's versioned scenario cards met that bar: each card carried context, tactic/technique labels, observables, recommended responses and reviewer provenance, and fed both SIEM rules and playbooks. That design also answered governance concerns that S. Sindiramutty *et al*. (2024) documented for smart-city cybersecurity – namely, operator-interpretable outputs. In practice, your curated cues – compressed inter-attempt timing, misordered stage traversals, eligibility mismatches and session correlations – were readily explainable to non-technical stakeholders and supported citizen-safe responses (rate limits, step-ups, paced slowdowns and clear denials).

A. Mohammed (2023) discussed audit-centric uses of AI in compliance. The present findings converged with both: soft, explainable controls outperformed black-box scoring in terms of auditability and user trust, and the memory artefacts created an audit-ready trail by design. Broader surveys such as F. Tao *et al*. (2021) also urged alignment with human decision-making; the achieved mean time-to-curation ≈ 22 min and curation acceptance rate ≈ 43% indicated that human-in-the-loop governance remained tractable. Two recent Ukrainian reviews further evidenced the policy relevance of AI-enabled defence. O.M. Lunhol (2024) catalogued AI-based methods and strategies in cybersecurity, stressing the need to balance capability with governance; Y.L. Vavryk & I.R. Opirskyy (2024) discussed "next-generation" AI for cybersecurity in national contexts. The present study complemented those perspectives by specifying how GenAI could be bounded for public platforms – descriptive hypotheses only, curated by experts, validated on twins – and by quantifying operational effects (for example, 39% False Positive Rate (FPR) reduction after introducing rolling baselines; 26% fewer escalation tickets per 10 k sessions). In that sense, the results supplied a procedural bridge between national-level strategy discussions and day-to-day SOC engineering.

Comparisons with adjacent empirical domains also proved informative. Y. Tovkun (2025) described cybercrime in digital employment, highlighting workflow misuse and identity friction. The authentication and document-entitlement findings here mirrored that pattern: small, legible irregularities in rhythm and role-document coherence flagged

misuse more reliably than signature-style rules, and re-authentication on device change separated benign context switches from opportunistic access with Positive Predictive Value ≈ 0.71 in twin tests. Meanwhile, provenance and reliability of AI outputs remained a live concern. K. Przystalski *et al*. (2025) showed that stylometry could detect LLM-generated text, underscoring the need for traceability. The present repository logged editor identity, timestamps and curation rationales, thereby addressing traceability and reducing the chance of unvetted prompts flowing into operations.

Where the present results diverged from prior applied work was in the minimal feature set required to reach operational value. Many studies relied on high-dimensional telemetry and deep anomaly detectors in industrial or critical-infrastructure contexts, whereas the e-government twins achieved high utility with per-step timing deltas, stage-edge traversals, eligibility states, and stable (hashed) session or device keys-fields that are privacy-preserving and inexpensive to instrument. On the governance side, audit-heavy, pre-deployment assurance has often been favoured; in contrast, the twin-only regimen enabled iterative learning without touching live systems and produced portable "immune-memory" artefacts that travelled across heterogeneous modules. Several limitations noted in the literature also manifested here: behavioural twins necessarily abstract infrastructure and vendors; concept drift remained a risk, hence prompts were versioned and model families recorded per artefact; and human effort concentrated in curation, with reviewer throughput indicating feasibility but still benefiting from peer review and a concise curation checklist.

Overall, the comparative picture was consistent: across independent strands – behaviour-level threat modelling, agentic-AI risk framing, digital-immune thinking, audit-first operations, and national reviews – the field moved towards explainable, governable defences. The present study advanced that trajectory by demonstrating that a twin-validated, memory-centric pipeline converted GenAI-generated hypotheses into ATT&CK-aligned detections and low-friction responses. Empirically, interpretable signals delivered measurable operational gains – FPR ↓ 39%, alert precision ↑ 23%, automated-cluster spikes ↓ 41% – while keeping privacy intact and governance ready. These comparisons supported the claim that bounded GenAI, when paired with digital twins and curated memory artefacts, offered a pragmatic route to strengthen everyday security of e-government platforms. Building on prior strands that called for future-proof e-governance security, explainable and auditable AI, and practical, testable AI-security threat models, this study constrained GenAI to produce interpretable scenarios, validated them on digital twins to avoid production risk, and embedded the outputs in a digital-immune loop of detection, reaction and memory across authentication, document access, registry and e-voting services.

Conclusions

Treating the model strictly as a producer of descriptive threat hypotheses, validating only behavioural signals on digital twins, and encoding outcomes as reusable immune-memory artefacts broadened adversarial coverage without disclosing exploits or touching live systems. Across identity, document, registry, and voting workflows, the most actionable indicators proved to be small, legible irregularities – compressed inter-attempt timing, misordered stage transitions, eligibility mismatches, and session-level correlations. These signals were inexpensive to instrument (timestamps, stage-graph logging, privacy-preserving fingerprints), traceable for audit, and mapped naturally to conservative, citizen-friendly responses (throttling, step-up authentication, pacing, clear denials). In each examined domain, such controls contained model-generated behaviours at the workflow layer while preserving availability and user trust.The scenario-to-memory pipeline operated as a bridge between research and operations.

Curated scenario cards – context, abstract tactic labels, observables, recommended responses, and provenance – were portable across services and readily integrated into detection-engineering backlogs and incident playbooks. Governance and privacy requirements were satisfied through human gatekeeping, minimal feature storage, explicit retention limits, and full reviewer attribution.The approach complemented, rather than replaced, penetration testing, formal verification, and compliance audits. Its distinctive value lay upstream: expanding the set of plausible behaviours worth monitoring and translating them into ATT&CK-aligned operational knowledge, yielding a richer catalogue of vetted detections and low-friction responses suitable for security operations centres. Language models occasionally over-generalised or proposed flows that conflicted with domain rules; expert curation mitigated this variance. Digital twins abstracted infrastructure and cryptographic proofs, so validation evidenced detectability and policy fitness rather than exploitability. Concept drift and model updates necessitated versioned prompts and model-family records. Fairness and user-impact checks remained necessary to ensure throttles and step-ups did not disproportionately affect particular cohorts or regions.

The work formalised a bounded, auditable role for GenAI in the public sector; operationalised a twin-only validation regime; defined a compact, portable immune-memory artefact linking hypothesis generation, SOC detections, and governance records; and isolated a minimal, cross-domain signal set (timing rhythms, stage-ordering anomalies, eligibility coherence, session correlations) that consistently yielded explainable, citizen-safe responses. Future priorities include an open, anonymised benchmark of curated scenarios for common government workflows; comparative studies that relate twin-validated signals to field telemetry; fairness and user-impact evaluation of throttling and step-up policies; cautious automation (semi-automatic Security Information and Event Management (SIEM) rule compilation, retrieval-augmented prompting under strict guardrails); and longitudinal deployments across election cycles and

peak-service periods to track immune-memory accumulation and threshold adaptation.

Acknowledgements

None.

Funding

The study received no funding.

Conflict of Interest

None.

References

[1] Adida, B. (2008). Helios: Web-based open-audit voting. In *Proceedings of the 17th USENIX security symposium* (pp. 335-348). Berkeley: USENIX Association.

[2] Alguliyev, R., Aliguliyev, R., & Yusifov, F. (2018). Role of social networks in e-government: Risks and security threats. *Online Journal of Communication and Media Technologies*, 8(4), 363-376. doi: 10.12973/ojcmt/3957.

[3] Al-Mushayt, O.S. (2019). Automating e-government services with artificial intelligence. *IEEE Access*, 7, 146821-146829. doi: 10.1109/ACCESS.2019.2946204.

[4] Arif, A., Khan, M.I., & Khan, A.R.A. (2024). An overview of cyber threats generated by AI. *International Journal of Multidisciplinary Sciences and Arts*, 3(4), 67-76. doi: 10.47709/ijmdsa.v3i4.4753.

[5] Basu, S. (2004). E-government and developing countries: An overview. *International Review of Law, Computers & Technology*, 18(1), 109-133. doi: 10.1080/13600860410001674779.

[6] Bécue, A., Praça, I., & Gama, J. (2021). Artificial intelligence, cyber-threats and Industry 4.0: Challenges and perspectives. *Artificial Intelligence Review*, 54, 3849-3886. doi: 10.1007/s10462-020-09942-2.

[7] Bodeau, D.J., McCollum, C.D., & Fox, D.B. (2018). *Cyber threat modeling: Survey, assessment, and representative framework*. McLean: The MITRE Corporation.

[8] De Santis, M., Esposito, C., & Mastroianni, M. (2025). Privacy risks in connected vehicles: Profiling threats and mitigation strategies. In O. Gervasi, B. Murgante, C. Garau, Y. Karaca, M.N. Faginas Lago, F. Scorza & A.C. Braga (Eds.), *Computational science and its applications – ICCSA 2025 workshops* (pp. 285-302). Cham: Springer. doi: 10.1007/978-3-031-97645-2_19.

[9] Diia. (n.d.). Retreived from https://diia.gov.ua/.

[10] George, A.S., George, A.S.H., & Baskar, T. (2023). Digitally immune systems: Building robust defences in the age of cyber threats. *Partners Universal International Innovation Journal*, 1(4), 155-172. doi: 10.5281/zenodo.8274514.

[11] Grosse, K., Dixit, P., Stark, E., Trinquier, V., Johansson, T., & Pinkas, B. (2024). Towards more practical threat models in artificial intelligence security. In *Proceedings of the 33rd USENIX security symposium* (4891-4908). Berkeley: USENIX Association.

[12] Guembe, B., Cáceres-Ortega, A., del Ser, J., Galar, M., Sanchis, A., & Sanz, R. (2022). The emerging threat of AI-driven cyber attacks: A review. *Applied Artificial Intelligence*, 36(1), article number 2037254. doi: 10.1080/08839514.2022.2037254.

[13] ISO 8601. (2019). *Date and time format*. Retrieved from https://www.iso.org/iso-8601-date-and-time-format.html.

[14] Jatkiewicz, P. (2025). Assessing cybersecurity methodologies: Integrating competitiveness factor for risk analysis and IT system design. *Expert Systems with Applications*, 296(D), article number 129220. doi: 10.1016/j.eswa.2025.129220.

[15] Kumar, R., Abdul Hamid, A., Ya'akub, N., Nyamasvisva, T., & Tiwari, R. (Eds.). (2025). *Leveraging futuristic machine learning and next-generational security for e-governance*. Hershey: IGI Global Scientific Publishing. doi: 10.4018/979-8-3693-7883-0.

[16] Lauer, T.W. (2004). The risk of e-voting. *Electronic Journal of e-Government*, 2(3), 167-186.

[17] Law of Ukraine No. 2297-VI "On Personal Data Protection". (2010, June). Retrieved from https://zakon.rada.gov.ua/laws/show/2297-17.

[18] Lunhol, O.M. (2024). Review of methods and strategies of cybersecurity using artificial intelligence. *Cybersecurity: Education, Science, Technique*, 1(25), 379-389. doi: 10.28925/2663-4023.2024.25.379389.

[19] Miah, M.N.I., Uddin, M.J., & Ahmed, M.W. (2025). AI-driven threat intelligence: Evaluating machine learning for real-time cyber threat sharing among U.S. national security agencies. *Journal of Computer Science and Technology Studies*, 7(8), 300-313. doi: 10.32996/jcsts.2025.7.8.34.

[20] Mohammed, A. (2023). Elevating cybersecurity audits: How AI is shaping compliance and threat detection. *Aitoz Multidisciplinary Review*, 2(1), 35-43.

[21] Moore, B.N. (2018). *Cyber threats in e-government*. (Doctoral dissertation, Northcentral University, San Diego, USA).

[22] Niarakis, A., *et al*. (2024). Immune digital twins for complex human pathologies: Applications, limitations, and challenges. *NPJ Systems Biology and Applications*, 10, article number 141. doi: 10.1038/s41540-024-00450-5.

[23] Pardue, H., Landry, J.P., & Yasinsac, A. (2011). E-voting risk assessment: A threat tree for direct recording electronic systems. *International Journal of Information Security and Privacy*, 5(3), 19-35. doi: 10.4018/jisp.2011070102.

[24] Przystalski, K., Argasiński, J.K., Grabska-Gradzińska, I., & Ochab, J.K. (2025). Stylometry recognizes human and LLM-generated texts in short samples. *Expert Systems with Applications*, 296(B), article number 129001. doi: 10.1016/j.eswa.2025.129001.

[25] Regulation of the European Parliament and of the Council No. 679 "On the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data and Repealing Directive 95/46/EC" (2016, April). Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj.
[26] Risnanto, S., Abd Rahim, Y., Mohd, O., Andinata, K., Effendi, A.R., & Perdana, R.S. (2021). E-voting: Security, threats and prevention. In *2021 15th international conference on telecommunication systems, services, and applications (TSSA)* (pp. 1-8). Piscataway: IEEE. doi: 10.1109/TSSA52866.2021.9768214.
[27] Schatz, M.C., & Phillippy, A.M. (2012). The rise of a digital immune system. *GigaScience*, 1(1), article number 2047-217X-1-4. doi: 10.1186/2047-217X-1-4.
[28] Sindiramutty, S.R., Tan, C.E., Lau, S.P., Thangaveloo, R., Gharib, A.H., Manchuri, A.R., Khan, N.A., Tee, W.J., & Muniandy, L. (2024). Explainable AI for cybersecurity. In M.M. Ghonge, N. Pradeep & N.Z. Jhanjhi (Eds.), *Advances in explainable AI applications for smart cities* (pp. 31-97). Hershey: IGI Global. doi: 10.4018/978-1-6684-6361-1.ch002.
[29] Tao, F., Akhtar, M.S., & Jiayuan, Z. (2021). The future of artificial intelligence in cybersecurity: A comprehensive survey. *EAI Endorsed Transactions on Creative Technologies*, 8(28), article number e3. doi: 10.4108/eai.7-7-2021.170285.
[30] Tovkun, Y. (2025). Cybercrime in the world of digital employment. *Collection of Scientific Papers "ΛΌГΟΣ"*, 225-231. doi: 10.36074/logos-13.12.2024.047.
[31] Vavryk, Y.L., & Opirskyy, I.R. (2024). Artificial intelligence: Cybersecurity of the new generation. *Ukrainian Scientific Journal of Information Security*, 30(2), 244-255. doi: 10.18372/2225-5036.30.19235.
[32] Weldemariam, K., Villafiorita, A., & Mattioli, A. (2007). Assessing procedural risks and threats in e-voting: Challenges and an approach. In A. Alkassar & M. Volkamer (Eds.), *E-voting and identity* (pp. 38-49). Berlin: Springer. doi: 10.1007/978-3-540-77493-8_4.
[33] Zambare, P., Thanikella, V.N., & Liu, Y. (2025). Securing agentic AI: Threat modeling and risk analysis for network monitoring agentic AI system. *ArXiv*. doi: 10.48550/arXiv.2508.10043.
[34] Zhao, J.J., & Zhao, S.Y. (2010). Opportunities and threats: A security assessment of state e-government websites. *Government Information Quarterly*, 27(1), 49-56. doi: 10.1016/j.giq.2009.07.004.

Застосування генеративних моделей штучного інтелекту для моделювання кіберзагроз у системах електронного урядування

Юлія Товкун

Аспірант
Харківський національний університет радіоелектроніки
61166, просп. Науки, 14, м. Харків, Україна
https://orcid.org/0009-0000-5916-2897

Анотація. Стрімка цифровізація перетворила державні платформи на об'єкти критичної інфраструктури, що потребують методів виявлення контекстних атак поза межами традиційних підходів. Метою було продемонструвати безпечну методику застосування генеративного штучного інтелекту для моделювання кіберзагроз у сервісах е-урядування з валідацією лише поведінкових сигналів на цифрових двійниках і кодуванням результатів у багаторазові артефакти «імунної пам'яті». Методика складалася з генерування описових «атакоподібних» сценаріїв, експертної курації, перевірки на мінімальних двійниках та формування детекцій і політик реагування. Отримано 170 гіпотез, 107 (62,9 %) відібрано після курації, 86 (80,4 % від відібраних) відтворено на двійниках. Для чотирьох кластерів зафіксовано метрики: точність 0,76-0,85, повнота 0,68–0,74, хибні спрацювання 0,4–1,2 %. Для входу точність/повнота 0,81/0,74; для «дрейфу повноважень» 0,85/0,69; для зондування реєстрів 0,79/0,71; для темпових сплесків у голосуванні 0,76–0,85. Реакції були малофрикційними: re-auth при зміні пристрою зменшила хибні відмови на 41 %; бюджети запитів і back-off скоротили підозрілі послідовності на 63 % без помітного впливу (< 0,2 %); «пейсинг» знизив кластерні спроби голосування на 58 %, а розсинхрон із перевіркою – на 46 %. Експлойти не створювалися; продуктивні системи не залучалися. Практична цінність – відтворюваний процес для команд кіберзахисту, операторів центрів оперативного управління безпекою і виборчих адміністрацій: перевірені сценарії трансформуються у правила моніторингу, політики помірного втручання (throttling, step-up, pacing, чіткі відмови) та версійовані артефакти знань, придатні до аудиту

Ключові слова: державні платформи; цифровий двійник; цифрова імунна система; електронне голосування; виборчі системи; політики реагування