

CyberNet: A Scenario-Based Web Serious Game for Everyday Cybersecurity Awareness

Design, Implementation, and a Within-Subjects Evaluation with 102 Participants

Mohammad Jaafar
Asinnah
<https://asinnah.com>

July 2026

Abstract

Cybersecurity breaches often result from human decisions rather than technical failures, yet traditional awareness training, delivered through static e-learning and compliance courses, rarely holds attention or changes behaviour. This paper designs, implements, and evaluates CyberNet, a scenario-based web serious game that encodes five everyday threats (phishing, password choice, voice phishing, removable-media baiting, and public Wi-Fi exposure) as short branched decisions with narrated outcomes and debriefs, grounded in the MDA, MOTENS, and Garriss frameworks. The five concept areas map onto a ten-item knowledge test, two items each. In a single-group, within-subjects, pre- and post-test study with $N = 102$ participants, mean accuracy rose from 54.80% to 77.75%, a gain of 22.94 percentage points (paired Cohen's $d = 1.07$, $p < 0.001$), and all five concepts reached family-wise significance under Holm–Bonferroni correction, with effects from $d = 0.39$ (vishing) to $d = 0.85$ (password). Confidently-incorrect answers fell from 39.8% to 20.3%, and moved into confident-correct rather than uncertainty. An eleven-item engagement survey scored high but did not correlate with the knowledge gain. The paper contributes a review of serious games for cybersecurity awareness, the deployed prototype released as an anonymous study instrument, empirical evidence from a within-subjects study, and a presentation-layer instrument that evaluates narration, music, and visual design as a distinct dimension. Limitations of the single-session design and three extensions (a control group, a delayed-retention follow-up, and a post-debrief behavioural probe) are discussed.

1 Introduction

Cybersecurity is a persistent concern for both individual users and organisations, yet human error remains one of the most common causes of security breaches. While organisations invest heavily in technical defences such as firewalls, encryption, and intrusion-detection systems, these measures often fail to address risky user behaviour. Research shows that mistakes such as clicking on phishing links, reusing weak passwords, and sharing sensitive information account for a large portion of successful cyberattacks [1], [2].

Traditional cybersecurity awareness programmes, delivered mainly through static e-learning modules or compliance-based courses, are often unengaging and fail to produce lasting behavioural change. In contrast, *serious games*, which are digital games designed for educational purposes beyond entertainment, offer a more interactive alternative. They allow users to explore realistic scenarios safely and to receive feedback in context, so that understanding develops through direct experience. Recent studies suggest that serious games can improve engagement, motivation, and security-related decision making [3], [4], [5].

Although many serious games have been created for education and awareness, there is limited evidence of their effectiveness in cybersecurity contexts. Some projects lack theoretical grounding, focusing more on entertainment than on measurable learning outcomes [6]. The challenge is to design serious games that are engaging as well as educationally effective, so that short-term knowledge gains are accompanied by more durable improvements in behaviour and decision making.

Objectives. The goal of this work is to design, implement, and evaluate a scenario-based serious game that improves everyday cybersecurity awareness and measurably shifts participants' recognition of common online threats. Specifically: (i) to review existing serious-game and gamified approaches to cybersecurity awareness and identify the design choices associated with measurable outcomes; (ii) to design and implement a web-based serious game that encodes five everyday attack vectors as branched interactive scenarios with typed outcomes and narrated debriefs; and (iii) to evaluate the prototype with human participants in a single-group pre- and post-test study, measuring both knowledge gain and self-reported engagement, including the distinct contribution of the presentation layer (narration, ambient music, and visual design).

Research questions. The study addresses three questions:

RQ1 (Design). Can a scenario-based web game encode five everyday cybersecurity threats as playable decisions whose branched outcomes teach the targeted learning objectives?

RQ2 (Learning). Does exposure to the prototype measurably improve participants' knowledge of the targeted concepts between the pre- and post-test, and does any improvement appear across all five concept areas or only on some?

RQ3 (Engagement and transfer). Do participants report the intervention as useful, easy to use, and likely to change their online behaviour? Does self-rated awareness correlate with measured knowledge, and do the presentation layers (narration, music, visual design) contribute to reported engagement?

Contributions. The paper contributes (a) a review of current serious games for cybersecurity awareness, mapped by learning goals, audiences, and evaluation outcomes; (b) a working scenario-based web prototype covering five everyday attack vectors, released as an anonymous browser-based study instrument; (c) empirical evidence from a single-group pre- and post-test user study with 102 analytic participants on how the prototype affects knowledge, confidence calibration, engagement, and self-rated awareness; and (d) a methodological extension to the engagement instrument: three presentation-layer items (narration, ambient music, visual design) separated from the standard Technology Acceptance Model items so that the cinematic surface can be evaluated explicitly.

The remainder of the paper is organised as follows. Section 2 reviews prior work. Section 3 presents the design, including the iteration that produced it. Section 4 describes the implementation. Section 5 sets out the study methodology, Section 6 reports the results, and Section 7 interprets them. Section 8 concludes with limitations and future work.

2 Related Work

The human element. A consistent theme in security failures is the role of human judgement under uncertainty. Hadlington reports that impulsivity, attitudes towards security, and broader digital habits correlate with risky behaviours online; the implication is that training must address motivational and cognitive antecedents rather than prescribe rules alone [1]. Dinet further highlights how social and cultural environments shape users' perceptions of risk and their willingness to follow protective practices in context [2]. Game-based approaches are attractive because they can safely simulate risk, encourage experimentation, and provide timely feedback aligned with how people learn and decide in real tasks.

Serious games in cybersecurity. Recent syntheses chart the evolution from early tabletop and simulation experiences to adaptive, data-informed interventions. Ng and Hasan review work from roughly 2014 to 2024 and observe growth in both general awareness games and more technical domains, along with a tendency to target learners who already possess some prior knowledge [3]. Moumouh et al. focus on professional settings and report promising outcomes but limited sectoral coverage and many small-sample evaluations [6]. For non-IT staff, Gwenhure and Rahayu identify practical considerations for workplace transfer and argue that design must reflect operational realities [7]. Alnajim et al. survey traditional training

alongside VR and AR [8], and Pahlavanpour and Gao map gamification within information-security programmes [9]. Collectively, these reviews suggest that while serious games can raise knowledge and intentions, stronger designs and deployment studies are needed to secure sustained behavioural change.

Gamification and behavioural change. Across workforce contexts, Bitrián et al. find that gamified e-training improves perceived usefulness and satisfaction which, in turn, raise self-efficacy; behaviourally, phishing click rates decrease following the intervention [4]. The Technology Acceptance Model (TAM) links perceived ease and usefulness to intentions and eventual use, offering constructs and items that can be embedded into evaluations of training tools [10]. Garris et al. describe an input–process–outcome loop for game learning in which feedback and debrief are critical to consolidating knowledge [11], while Gee argues that effective games promote situated meaning and transfer [12]. In security-specific contexts, engagement gains are common [5], and participatory designs appear increasingly promising [13]. Controlled evidence grounded in the Theory of Planned Behaviour shows that security games can shift attitudes, perceived behavioural control, intentions, and self-reported behaviour relative to non-security game baselines [14]. Topic-focused comparisons of education-based and game-based approaches report benefits alongside the need for continual recalibration as threat tactics evolve [15], and model-driven adaptive systems such as CyberHero show how tailoring difficulty and feedback sustains engagement [16].

Pedagogical and design frameworks. The Mechanics–Dynamics–Aesthetics (MDA) lens maps game mechanics to the dynamics they produce and the aesthetic experience they aim to evoke, and offers a shared vocabulary across pedagogy and design [17]. Constructivist approaches emphasise authentic tasks, reflection, and active meaning-making [18]. In cyber-specific design, MOTENS ties threat scenarios to formative feedback and observable outcomes [19]. Recent work also advances staged training models for awareness programmes [20] and domain-general mappings between learning goals and game mechanics [21]. Design lenses can target deceptive interfaces, as in the *Trickery* project [22], and tabletop co-operative formats such as tCOFELET scaffold reflection and shared mental models [23].

Evaluation methods. Studies typically operationalise effectiveness through knowledge and skill tests, behavioural proxies (simulated phishing click rates, decision logs, telemetry), affective and cognitive constructs (self-efficacy, TAM items, intention), and retention or transfer measures. Labunets et al. show how perception measures complement task yield [24]; CyberCIEGE produces event logs that support after-action review [25], [26]. Most training studies use short-term post tests; fewer include delayed retention or operational outcomes.

Empirical studies and modalities. Phishing-focused interventions report improved recognition and reduced susceptibility [27]. Social-engineering games have been used to elicit decision rationales that reveal misconceptions [28], and developer-facing games show awareness gains for secure practices [29]. Card games such as Control-Alt-Hack remain influential for engagement and reflection in group settings [30], [31]; competitive gamification can sustain interest but must be aligned with learning goals [32]; cyber ranges and CTF-as-a-service provide realism and time pressure [33]. Le-Nye et al. compared traditional courses and interactive phishing games across 115 participants [15], and a cyber-physical systems game surfaced distinct decision heuristics across stakeholder groups [34]. Other modalities include escape rooms [35], culturally grounded designs such as CyberMoraba [36], and immersive, data-rich XR platforms [37]. Overall, effect sizes are strongest for knowledge and intentions; sustained behaviour change is more variable without boosters or organisational support.

Gaps. Three gaps recur across the reviews. External validity is limited by convenience samples and classroom settings [3], [6], [7]. Theory-driven design and measurement are inconsistently operationalised, and more transparent mappings from learning objectives to mechanics and metrics are needed [17], [18], [19], [21]. Finally, effects taper without reinforcement, and emerging immersive platforms raise privacy

and ethics questions for telemetry [37], [38]. Building on these insights, the present work adopts a focused, theory-informed design with clear learning objectives, immediate explanatory feedback, and an evaluation that pairs a knowledge instrument with TAM-grounded engagement constructs [10], [11], [14], [15].

3 Design

3.1 Design iteration: from Ludii to a scenario web game

An earlier, exploratory iteration of the prototype was developed in the Ludii general-game system as a partial-information board game. Ludii was chosen because its ludeme language supports declarative modelling of asymmetric-information games [39] and because it ships with a general-purpose Monte Carlo Tree Search baseline. Informal playtests revealed two recurrent issues. First, the chess-like abstraction rewarded strategic inference over recognition of specific cybersecurity concepts, and participants reported that the mapping from board state to real threats was hard to recover without extensive priming. Second, the literature reviewed above suggests that scenario-based designs are a better fit than abstract strategy games for behavioural-awareness outcomes [13], [19], [21], [27].

Motivated by both the playtest evidence and the learning objectives below, the prototype was re-scoped to a scenario-based web artefact in which each scenario is a short, recognisable everyday moment rather than a move on an abstract board. Section 7.3 returns to this pivot as a finding in its own right.

3.2 Learning objectives

The prototype targets recognition of five everyday attack vectors, each stated at the identify–explain–choose level and tied to a single scenario:

1. **Phishing signal recognition** — identify the cues that a routine-looking work email is in fact a lure (sender domain mismatch, pressure framing, suspicious link).
2. **Password strength judgement** — distinguish candidates that resemble strength (leet substitutions, name-plus-year) from candidates whose structure resists common attacks (passphrases, random strings paired with a manager).
3. **Caller-identity verification** — recognise the typical vishing script and the appropriate response of independent callback through a verified number.
4. **Physical-media baiting** — recognise that a found storage device is a plausible social-engineering vector and that the safe response is to hand it to IT rather than inspect it personally.
5. **Public Wi-Fi trust model** — recognise that open hotspots, captive-portal networks, and cellular tethering carry distinct exposure profiles, with tethering the safest default.

Each objective maps onto two items of the ten-item knowledge instrument (Section 5.3) and onto the debrief beat of the corresponding scenario.

3.3 Design principles

Six principles drawn from the frameworks above shape the scenarios. Scene-level feedback loops [11] pair every decision with a typed outcome beat and a concept-level debrief, so the causal link from choice to consequence is not left implicit. Progressive difficulty runs across the sequence: early scenarios reward surface-signal recognition, later scenarios require structural reasoning about trust and network sharing. Narrative immersion anchors each scenario in a plausible everyday moment (an inbox email, a phone ringing, a USB on the floor). Goal clarity [17] is maintained by stripping each decision scene to the workspace the player needs. Agency is bounded to three or four options per decision, avoiding both false dichotomies and choice paralysis. An explicit post-play debrief [19] closes each scenario with a narrated takeaway and a longer lesson paragraph, consistent with constructivist accounts of experience-plus-reflection [18]. The presentation layer (narrated voice, ambient music, consistent SOC-analyst visuals) is treated as part of the design and evaluated explicitly in the engagement survey.

3.4 The five scenarios

Each scenario is a short directed graph of typed scenes with three or four options at the decision point.

Phishing: IT helpdesk. The participant is shown an inbox with a single unread message purporting to come from an IT helpdesk and asking them to confirm account details through a link. The email exposes inspectable hotspots on the sender address, the subject line, and the link target; clicking a hotspot surfaces a short explanation of the associated red flag inline. The decision offers three branches: click the link, report the message as phishing, or delete it without reporting. The embedded link is itself a trap: clicking it advances directly to the compromised outcome, so an accidental click has the same consequence it would on a real inbox.

Password: pick your password. The participant is shown an account-settings prompt indicating that their current password has expired. Four candidate replacements are offered: a leet-substituted dictionary word (Tr0ub4d0r&3), a three-word passphrase (sunset-piano-bicycle), a short random string (aX9\$kL2!), and a memorable personal pattern (Jessica1987!). An earlier version asked the participant to type a password directly; pilot observations indicated that participants consistently typed strong passwords and therefore never encountered the weak-password outcomes the scenario is designed to surface. The multiple-choice redesign ensures every participant is exposed to the full outcome space. The four options route deterministically to three outcomes covering passphrase strength, dictionary-plus-substitution weakness, and name-plus-year predictability.

Vishing: support-desk scam. The participant receives a call from a caller who identifies himself as a member of a well-known vendor’s support team. The scenario is framed inside a phone interface with three phases: a ringing phase with a looping ringtone; a connected phase in which three scripted caller lines are played through a distinct voice with typed subtitles synchronised to audio playback; and a deciding phase that presents three branches (comply with the request, hang up and call back through a known number, or hang up without callback). The implementation uses the participant’s first click both to unlock browser audio and to accept the call.

USB drop. The participant is shown a USB stick found in a workplace setting, carrying a handwritten paper label reading “HR – Bonuses 2026 – CONFIDENTIAL”. Three branches are offered: plug it into the participant’s own laptop, leave it in a shared break room, or hand it to the IT department without inspecting it. The outcomes cover the plausible attack mechanics associated with dropped storage devices [40], including keystroke-injection hardware, payloaded autorun, and credential-harvesting documents, together with the constructive option of returning the device to IT.

Public Wi-Fi. The participant is shown a mobile Wi-Fi settings screen with three networks available: an open hotspot configured as an evil twin, a legitimate café network requiring a captive-portal sign-in, and an explicit option to use cellular tethering. The scenario teaches that open networks expose traffic to passive observers with no barrier to entry [41], that shared-key and captive-portal networks still mix participants on a common segment, and that cellular tethering separates the device from third-party infrastructure.

4 Implementation

The prototype is a web application, publicly available at <https://cybernet-awareness.vercel.app/>.

4.1 System architecture

The client is built on Next.js, React, and TypeScript. Persistence is provided by a managed PostgreSQL service (Supabase) accessed over its REST layer with row-level security enabled. Narration is pre-generated through the ElevenLabs text-to-speech service and shipped as static audio assets, so no external API is contacted during a participant’s session (Figure 1).

Each choice was made against a short list of alternatives. A web stack was preferred over a native client because the recruitment model relies on a single shareable link, and a browser artefact lowers participation cost to a click. Next.js with the App Router was preferred over a plain React setup because

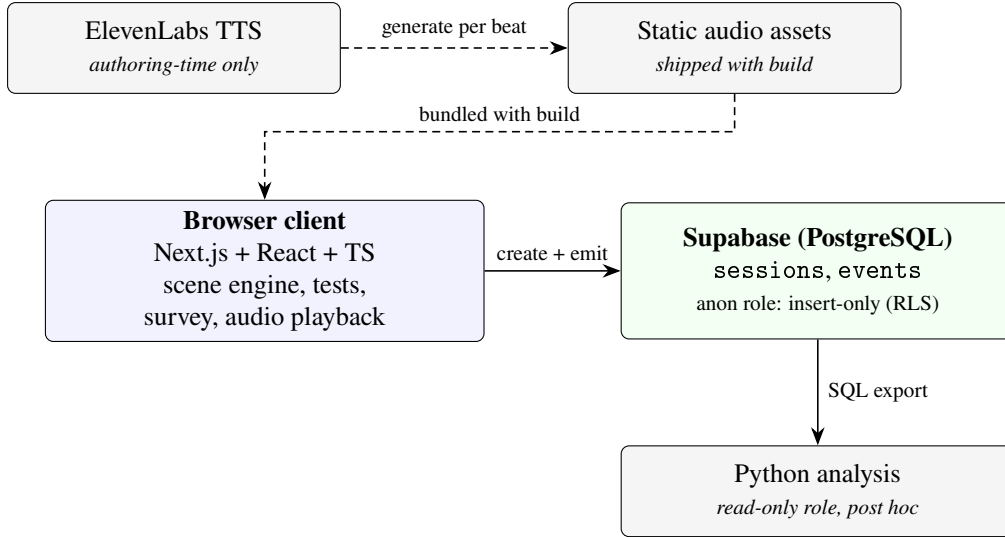


Figure 1: CyberNet runtime architecture. Solid arrows are runtime flows during a participant session; dashed arrows are authoring-time flows. The browser-side client has no runtime dependency on the external text-to-speech service.

the routed page model maps directly onto the linear participant flow. Supabase was preferred over self-hosted PostgreSQL and over a NoSQL alternative because the relational model fits the two-table schema in Section 4.5, row-level security gives the anonymous client an insert-only posture by configuration, and European-region hosting aligns with the consent commitments. Deployment on Vercel removes operational responsibility for a study-period outage.

4.2 Scenario engine

The scenarios are expressed as small directed graphs of typed scenes. Four scene types are used: *stimulus* scenes present narrative text and advance automatically; *decision* scenes branch on the participant’s choice; *outcome* scenes narrate the consequence of a chosen branch; and *debrief* scenes consolidate the lesson with a takeaway line and a longer explanation. A fifth *quiz* type is defined for a single-question recap but is not used in the present study. A runtime scene runner walks the graph, handing the current scene to a rendering surface and advancing on participant input.

A directed graph of typed scenes was chosen over two alternatives. Authoring tools such as Twine or Ink offer a polished writing experience but couple the branching content to their own runtimes, which would complicate the in-application logging the analysis relies on; the typed-scene approach keeps decisions, outcomes, and debriefs in the same source tree as the rest of the application and lets the event logger observe scene transitions directly. A purely linear, scripted flow was rejected because it would collapse the difference between *making a decision* and *seeing its consequence*, on which the calibration analysis depends.

Custom visual surfaces are attached to scenarios through a small registry pattern: an email card with inspectable hotspots for phishing, an account-settings form for the password scenario, a phone frame with a three-phase call machine for vishing, a physical-object surface for USB, and a mobile-device settings screen for public Wi-Fi. The surfaces are cosmetic; the scene graph is the single source of truth for decision content.

4.3 Audio pipeline

Narration is pre-generated. A development-time script walks the scenario data, splits each narrated passage into sentence-level beats, and calls the text-to-speech API once per beat. The resulting audio files are written to deterministic paths derived from the beat identifier and served as static assets, so scenario

playback does not depend on network conditions or third-party availability at the time of participation. Pre-generation was preferred over a runtime call because a live API call adds latency tied to the participant’s network, because the per-beat cost is paid once rather than per participant, and because deterministic asset paths make the rendered scenario reproducible even if the upstream service updates its model. Two voices are used: a cinematic narrator for the introductory sequence, outcome beats, and debriefs, and a distinct caller voice for the vishing character. Narrator and background-music channels are independent, and volume adjustments propagate to every audio element currently playing.

4.4 Study flow

The participant route is fixed and is shown with approximate timings in Figure 2. A random session identifier is generated on consent, stored in the browser, and attached to every subsequent event; every gated page redirects to the landing route when no session identifier is present. A briefing screen between consent and the pre-test outlines the four-step arc the participant is about to follow, so that the post-test at the end is not encountered as an unexpected step.

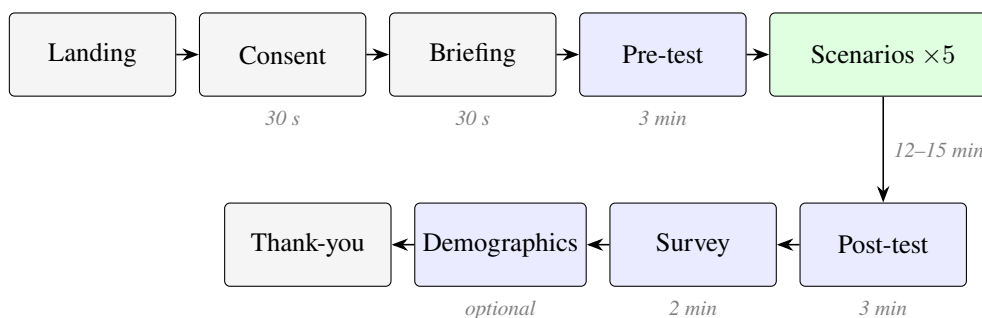


Figure 2: Participant route through CyberNet. Framing screens (grey) bookend the session; the five scenarios (green) carry the intervention; the pre- and post-tests and self-report instruments (blue) capture the dependent measures.

4.5 Data capture

Two database tables are used. A *sessions* table carries one row per participant, with a session identifier, a creation timestamp, and an optional completion timestamp. An *events* table carries one row per meaningful participant action with a session reference, a timestamp, an event type, and a JSON payload. The event vocabulary covers consent, pre-test item responses and summary, scene-level choices, post-test item responses and summary, survey responses, demographics, and session-end. Row-level security is enabled on both tables: the anonymous client key grants insert-only access, and read access is gated behind a read-only role used for analysis. Event writes are non-blocking so that a transient network failure does not interrupt the session. The application captures no names, no email addresses, and no IP addresses; session records carry only a random identifier, are stored in a European region, and are subject to the two-year retention horizon declared on the consent page.

4.6 Accessibility and pilot

Option buttons, confidence ratings, and Likert responses are marked up as radio groups so screen readers expose the selection state correctly. Focus outlines remain visible on both light and dark button fills. The `prefers-reduced-motion` preference collapses page-level animations, and the entire participant route is operable with the keyboard alone, including a keyboard equivalent for the click-to-advance narrative surface. A small internal pilot verified the end-to-end flow and surfaced three adjustments: the password scenario moved from free-text entry to the multiple-choice design, a briefing screen was added between consent and the pre-test, and accessibility gaps (missing focus indicators on amber fills, inconsistent screen-reader cues on Likert items) were resolved before broader recruitment.

5 Study Methodology

5.1 Research design

The evaluation follows a mixed-methods design: a single-group, within-subjects, pre- and post-test quasi-experimental study in which each participant serves as their own control, accompanied by a post-intervention engagement survey and an optional demographic form. Single-group pre/post is the modal design in recent scenario-based awareness work [4], [14], [15] and is appropriate when the research question concerns within-participant change rather than relative efficacy against a rival intervention. Its main cost is that observed gains cannot be cleanly attributed to the intervention in the presence of a test-retest effect; this is addressed under threats to validity.

5.2 Participants

The target population is adults with everyday digital exposure and no required cybersecurity background. Recruitment was conducted through convenience sampling via student communities, Discord servers, and the author’s personal network, with no compensation offered. A target of at least fifty completed sessions was set, informed by the sample sizes reported in comparable studies [15], [42]; larger samples were accepted. Inclusion criteria were age eighteen or older, adequate English or Arabic reading level, and browser access. No participant was excluded on prior training; training level was captured in the demographic form and used as a covariate.

5.3 Instruments

Informed consent. The consent page presents anonymity, a two-year retention horizon, an ethics contact address, and the fact that anonymous sessions cannot be traced for post-hoc withdrawal. Participation is voluntary and participants may stop at any time by closing the tab.

Pre-test (ten items). Ten items cover the five concept areas, two per concept. Each item has four options, each written as something a participant might plausibly do or believe, so the correct answer is not recognisable from its wording alone. Most items are framed as a first-person decision; the remainder use an evaluative framing where the concept calls for reasoning about how an attack or defence works. Correct-answer positions are balanced across the four slots. Each item also collects an optional binary confidence marker (*not sure* or *sure*).

Post-test. The post-test uses the same ten items presented in reverse order. Parallel-form isomorphic items were considered and rejected because they would dilute the paired comparison.

Engagement survey (eleven items, nine constructs). Eight items cover six constructs drawn from the Technology Acceptance Model and the gamification-engagement literature [4], [5], [10], [14]: ease of use, perceived usefulness, perceived learning, self-efficacy, engagement (two items), and behavioural intention (two items). Three further items target the presentation layer explicitly: narration helpfulness, music contribution, and visual engagement. All items are rated on a five-point scale.

Demographics (twelve items). Age range, gender, current role, field, primary language, prior cybersecurity training, password-manager use, two-factor-authentication use, prior phishing exposure, public Wi-Fi frequency, risk-tolerance self-rating, and an overall self-rated awareness rating. Demographics are collected after the intervention, and each item is skippable, so that questions about the participant’s own expertise do not prime the pre-test.

An in-game event log records every decision, every outcome reached, and scene-level timestamps.

5.4 Procedure

Each participant completes the study in a single session of approximately twenty minutes: informed consent, a short briefing, the ten-item pre-test, a short introductory sequence that sets the fictional

Tuesday-morning framing, the five scenarios in order, the ten-item post-test (items reversed), the eleven-item engagement survey, the optional demographic form, and a thank-you screen. The scenarios take approximately twelve to fifteen minutes; the knowledge test adds about three minutes at each end.

5.5 Analysis plan

The primary analysis is a paired-samples *t*-test comparing pre-test and post-test overall accuracy, with a Wilcoxon signed-rank test where the normality assumption fails; effect sizes are reported as Cohen's *d*. A per-concept analysis runs the same paired test within each of the five concept areas (two items per concept) under Holm–Bonferroni correction. Covariate analysis stratifies on prior-training level and self-rated awareness. Engagement items are reported descriptively at the construct level, with the composite engagement score correlated against the pre/post delta; the three presentation items are reported separately. Calibration is assessed through the rate of confident-wrong responses, compared pre- and post-test. Analysis is scripted in Python (pandas, scipy, statsmodels) against a read-only export of the event log.

5.6 Ethical considerations

The study involves no deception, no exposure to real credentials, and no collection of sensitive personal data. Data are pseudonymised by design: the application captures no names, email addresses, or IP addresses, and session records carry only a random identifier. All records are stored in a European region.

6 Results

All numerical results were produced by a single analysis script run against the participant event log.

6.1 Participant flow and sample

Recruitment over a four-week window produced 142 session records on the live build. The analytic sample, $N = 102$, applies the criterion that both the pre-test and the post-test must have been completed in full (ten items each). Every result below is reported against this denominator. Within the pool, the engagement survey was completed in full by all 102 participants, the demographic form was returned by all 102 (with item-level skipping yielding between 90 and 98 responses per item), and at least one in-scenario choice event was captured for all 102.

The demographic spread is summarised in Table 1. Most respondents fall in the 18–24 ($n = 50$) or 25–34 ($n = 44$) age bands and identify as a student, an IT professional, or a non-IT professional. Prior cybersecurity training is mixed: the modal category is “None” ($n = 36$), with a comparable share reporting a formal course or module ($n = 31$). Self-rated awareness clusters at 3 and 4 on a five-point scale. The deployed build was localised in English and Arabic; categorical responses were normalised to a single English label scheme, so the counts pool the two language groups.

6.2 Overall pre- and post-test knowledge

Across the 102 analytic participants, mean pre-test accuracy was 54.80% (SD 19.69) and mean post-test accuracy was 77.75% (SD 19.85), a mean within-subject improvement of 22.94 percentage points (SD 21.37). A paired-samples *t*-test rejected the null hypothesis of no change, $t(101) = 10.84$, $p < 0.001$, with a paired Cohen's $d = 1.07$ (Table 2). Because a Shapiro–Wilk test on the difference scores returned $p = 0.018$, a Wilcoxon signed-rank test was run as a non-parametric companion and reached the same conclusion, $W = 196.5$, $p < 0.001$.

Table 1: Demographic summary across the analytic sample. Counts are responses to the optional post-intervention form.

Item	Responses	Most-common categories (count)
Age	98	18–24 (50), 25–34 (44), Under 18 (4)
Gender	98	Male (52), Female (39), Prefer not to say (7)
Role	98	Student (31), Non-IT professional (28), IT professional (22), Other (17)
Prior training	98	None (36), Formal course or module (31), Informal (articles, videos, work emails) (20), Prefer not to say (9)
Self-rated awareness	96	4.0 (36), 3.0 (26), 5.0 (16), 2.0 (13)
Password manager	91	No (47), Not sure what that means (22), Yes (14), Prefer not to say (8)
Two-factor auth	90	Most or all accounts (61), Some accounts (18), None (9), Not sure what that means (2)
Public Wi-Fi use	90	Occasionally (47), Regularly (30), Never (13)
Prior victim	90	No (40), Yes (26), Not sure (20), Prefer not to say (4)
Risk tolerance	90	3.0 (42), 2.0 (24), 4.0 (16), 5.0 (4)
Primary language	91	English (70), Arabic (21)

Table 2: Pre- and post-test accuracy across $N = 102$ analytic participants.

	Pre	Post	Δ
Mean accuracy (%)	54.80	77.75	+22.94
SD (%)	19.69	19.85	21.37
Median (%)	50.0	80.0	+20.0

Paired $t(101) = 10.84, p < 0.001$; Cohen’s $d = 1.07$.
Wilcoxon signed-rank $W = 196.5, p < 0.001$; Shapiro–Wilk on diffs $p = 0.018$.

6.3 Per-concept knowledge

Each of the five concept areas was probed by two pre-test items and two post-test items. A paired t -test was run within each concept, and the family of five comparisons was controlled using the Holm–Bonferroni procedure (Table 3, Figure 3). The two-item-per-concept instrument supports only directional interpretation: the per-concept analysis is exploratory, a description of where the aggregate effect concentrates rather than a confirmatory test at any single concept.

Table 3: Per-concept paired analysis ($N = 102$ within each concept). Raw p -values are reported alongside Holm-corrected family-wise values; concept-level effect sizes are exploratory.

Concept	n	Pre (%)	Post (%)	Δ (%)	t	p	p_{Holm}	d
Phishing	102	55.4	77.9	+22.5	5.95	<0.001	<0.001	+0.59
Password	102	38.2	73.0	+34.8	8.59	<0.001	<0.001	+0.85
Vishing	102	61.8	79.4	+17.6	3.93	<0.001	<0.001	+0.39
USB drop	102	58.8	79.9	+21.1	5.16	<0.001	<0.001	+0.51
Public Wi-Fi	102	59.8	78.4	+18.6	4.81	<0.001	<0.001	+0.48

Read directionally, the largest concept-level gain is on password (+34.8 percentage points, $d = 0.85$) from the lowest baseline (38.2% pre-test accuracy), and the smallest is on vishing (+17.6 points, $d = 0.39$) from the highest baseline (61.8%). The remaining three concepts sit in between: phishing $d = 0.59$, USB drop $d = 0.51$, public Wi-Fi $d = 0.48$. No concept shows a non-significant gain at the family-wise level, so the aggregate effect is not driven by any single area.

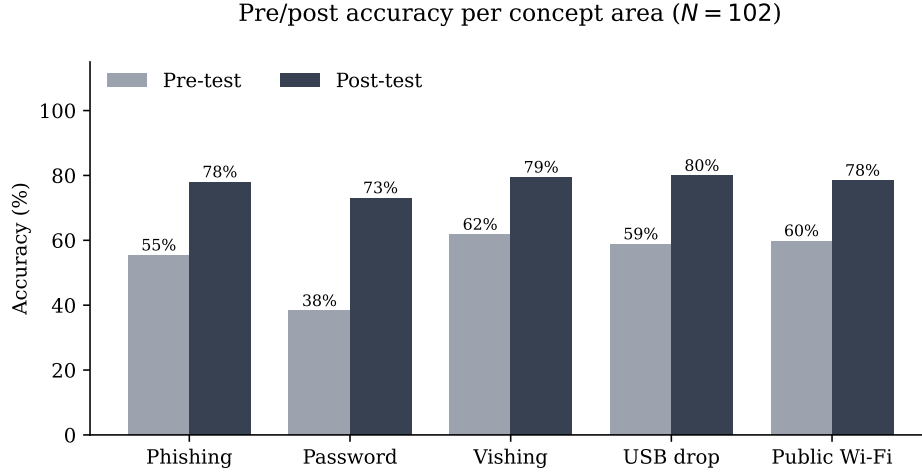


Figure 3: Mean pre- and post-test accuracy by concept area, $N = 102$.

6.4 Gain by demographic covariate

Tables 4 and 5 stratify the within-subject gain by prior-training level and by self-rated awareness. Because these items are optional, 98 of the 102 participants reported a prior-training level and 96 a self-rated awareness level; the others are excluded only from the corresponding stratification.

Table 4: Mean pre/post knowledge gain stratified by prior cybersecurity training (98 of 102 participants).

Prior training	n	Mean gain (pp)	SD
None	36	+27.78	18.53
Informal (articles, videos)	20	+24.50	21.14
Formal course or module	31	+21.94	21.82
Certification	2	+20.00	14.14
Prefer not to say	9	+22.22	14.81

Table 5: Mean pre/post knowledge gain stratified by self-rated cybersecurity awareness (96 of 102 participants).

Self-rated awareness	n	Mean gain (pp)	SD
1 (lowest)	5	+24.00	8.94
2	13	+23.85	15.57
3	26	+29.62	22.54
4	36	+22.50	19.91
5 (highest)	16	+22.50	19.83

The gain is broadly distributed across both stratifications. By training level, every category shows a positive change; the “None” group ($n = 36$) shows the largest mean gain at +27.8 points and the “Certification” category the smallest at +20.0, though the latter rests on two responses and is read indicatively only. The pattern is consistent with the prototype being calibrated to non-experts: participants with no formal training had the most room to improve, and improved the most. By self-rated awareness, all five rating levels sit in the +22.5 to +29.6 point band, with no monotone trend across the scale.

6.5 Calibration: confidence and correctness

The binary confidence marker on each item permits a calibration analysis: how often participants reported confidence on items they answered incorrectly, before and after the intervention. The share of items in

the *confident-wrong* cell fell from 39.8% on the pre-test to 20.3% on the post-test (Table 6, Figure 4). A paired t -test on the per-participant confident-wrong rate yields $t(101) = -9.39, p < 0.001, d = -0.93$. The complementary movement is a rise in the *confident-correct* cell from 49.9% to 72.9%; the *unsure* cells move only marginally. Participants do not become more uncertain across the board: items on which they were confidently wrong before tend to migrate into the confidently-correct cell after, rather than into either unsure cell.

Table 6: Distribution of items across the four confidence-by-correctness cells, before and after the intervention.

	Confident–Correct	Confident–Wrong	Unsure–Correct	Unsure–Wrong
Pre	49.9%	39.8%	4.8%	5.3%
Post	72.9%	20.3%	4.6%	1.8%
Confident-wrong rate: paired $t(101) = -9.39, p < 0.001$; Cohen’s $d = -0.93$.				

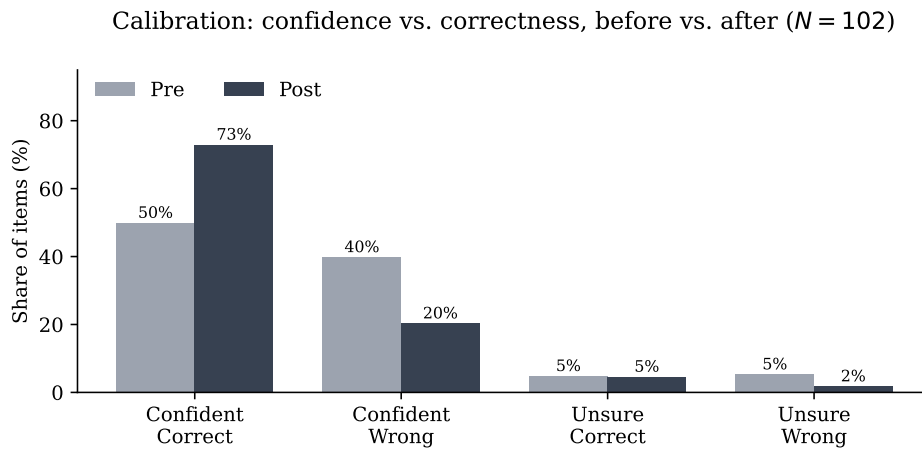


Figure 4: Item-level distribution across the confident-correct, confident-wrong, unsure-correct, and unsure-wrong cells, before and after the intervention.

6.6 Engagement and perception

Table 7 reports per-construct means together with the top-two-box rate (the share of responses scoring 4 or 5). The TAM constructs score consistently high, with perceived usefulness (mean 4.48, top-two-box 91.2%), perceived learning (4.42, 89.2%), and behavioural intention (4.41, 86.3%) at the top (Figure 5). The three presentation items show a wider spread: visual engagement is well received (4.25, 81.4%), narration helpfulness sits a step lower (3.95, 74.5%), and the music contribution item draws the broadest dispersion (3.64, 61.8%).

A composite engagement score (the mean of the six TAM constructs) was correlated against the per-participant knowledge delta. The correlation is small and not statistically reliable, $r = -0.05, p = 0.65$ (Pearson); the Spearman counterpart is comparable, $\rho = -0.14, p = 0.17$. The data are compatible with no monotonic relationship between how favourably a participant rated the experience and how much measured knowledge they gained.

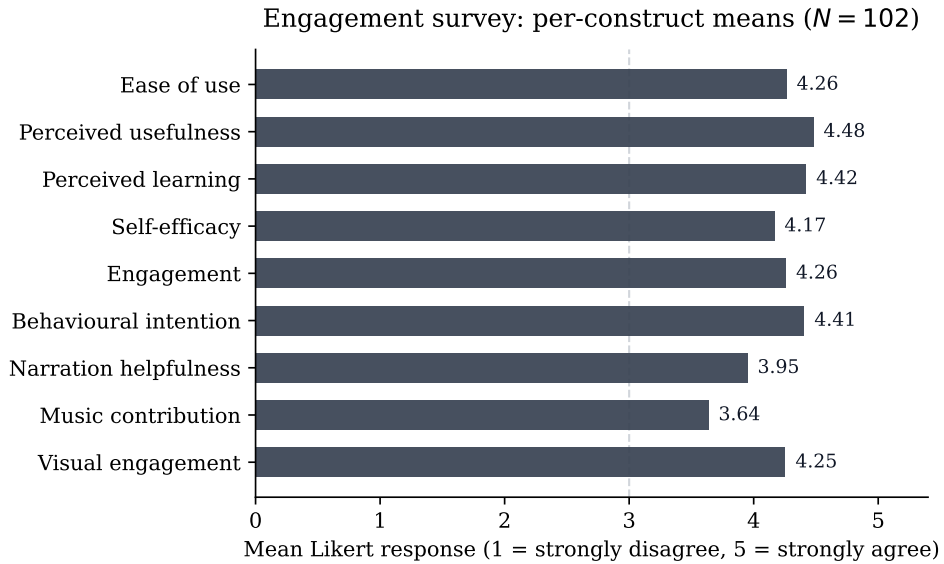
6.7 In-game decisions per scenario

The scene-level event log records which option each participant selected at the decision point of each scenario. A safe-decision rate is defined per scenario as the share of participants whose chosen option leads to a non-compromised outcome (Table 8, Figure 6). The rank-order runs from vishing (100%), through phishing (94.1%), public Wi-Fi (82.4%), USB drop (73.5%), to password (53.9%).

Table 7: Engagement survey descriptives across the nine constructs, $N = 102$.

Construct	n	Mean	SD	Median	Top-2 box
Ease of use	102	4.26	1.08	5.0	85.3%
Perceived usefulness	102	4.48	0.73	5.0	91.2%
Perceived learning	102	4.42	0.81	5.0	89.2%
Self-efficacy	102	4.17	1.14	4.5	82.4%
Engagement	102	4.26	0.75	4.5	84.8%
Behavioural intention	102	4.41	0.70	4.5	86.3%
Narration helpfulness	102	3.95	1.08	4.0	74.5%
Music contribution	102	3.64	1.45	4.0	61.8%
Visual engagement	102	4.25	1.01	5.0	81.4%

Composite TAM engagement (mean of 6 TAM constructs) vs. pre/post Δ : Pearson $r = -0.05$, $p = 0.647$; Spearman $\rho = -0.14$, $p = 0.169$.

**Figure 5:** Per-construct mean Likert response on the engagement survey. The dashed line marks the scale mid-point.**Table 8:** Per-scenario safe-decision rate across the $N = 102$ analytic sample.

Scenario	Safe-decision rate	Safe / total
Phishing (IT helpdesk)	94.1%	96 / 102
Password choice	53.9%	55 / 102
Vishing call	100.0%	102 / 102
USB drop	73.5%	75 / 102
Public Wi-Fi	82.4%	84 / 102

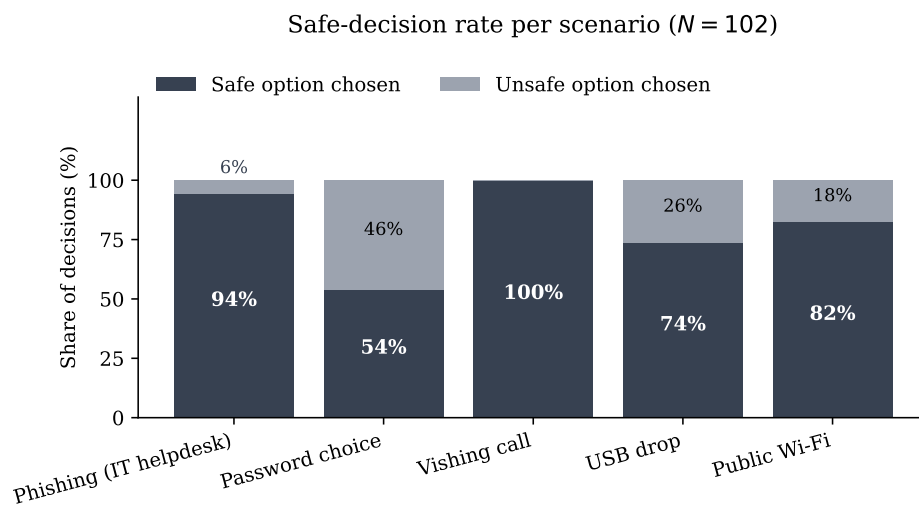


Figure 6: Share of decisions choosing the safe option, per scenario.

7 Discussion

7.1 Research questions

RQ1 (Design). The prototype encodes five everyday cybersecurity threats as playable branched decisions in which every choice produces a typed outcome and a narrated debrief. Each scenario reached every analytic participant and terminated in a named outcome, and the engagement survey reports broad endorsement of the design’s perceived realism. Whether the branched outcomes *teach* the targeted objectives is answered by the per-concept gains: every concept area reaches family-wise significance, so the design is not merely playable but transmits the targeted learning.

RQ2 (Learning). The knowledge instrument measures recognition of everyday cybersecurity cues rather than deep procedural expertise. The within-subject gain is large in the aggregate (+22.94 points, $d = 1.07$) and reaches family-wise significance across all five concept areas. The calibration analysis adds a complementary result: the share of items on which a participant is both confident and wrong falls from 39.8% to 20.3%, and those items move into the confident-correct cell rather than into uncertainty. RQ2 is answered in the affirmative at the aggregate level, subject to the established caveat of the single-group design [14]: a portion of the change may reflect a test-retest effect that the design cannot separate from the intervention. The concept-level pattern is read directionally; with two items per concept it describes where the aggregate effect concentrates and does not support a separate confirmatory claim at any single concept.

RQ3 (Engagement and transfer). Self-reported engagement is high across the TAM constructs (means 4.17–4.48, top-two-box 82.4%–91.2%), and behavioural intention is among the highest-scoring items. The presentation layer is more nuanced: visual engagement is endorsed, narration helpfulness is widely but not unanimously endorsed, and the music contribution is the lowest-rated construct. The correlation between composite engagement and the knowledge delta is small and not reliable, and gain by self-rated awareness sits within a narrow band across all five rating levels, so a participant’s own awareness rating does not predict the size of their improvement. RQ3 is therefore answered partially: participants endorse the intervention as useful, easy to use, and likely to influence their behaviour, but neither composite engagement nor self-rated awareness predicts the knowledge delta at the participant level. A comparable dissociation is reported in the gamification literature [5].

7.2 The knowledge–behaviour gap

The clearest within-session indication of a gap between stated knowledge and observed behaviour is the password scenario. Post-test accuracy on the two password items reaches 73.0%, but the in-scenario safe-decision rate at the moment the participant has to pick a password is only 53.9%: roughly nineteen points lower, on the same participants, within the same session. The order of operations puts the decision before the debrief by design, and the gap reflects pre-debrief intuition surviving into the choice rather than the debrief failing to teach. The pattern fits the human-factors literature in which stated knowledge does not translate into behaviour at a one-to-one rate [1], [2]. The implication for design is concrete: an intervention that hopes to shift behaviour as well as knowledge needs to test choice *after* the corresponding debrief, not before.

The rank-order of safe-decision rates is also informative. Scenarios that reward recognising a single high-salience cue (a suspicious caller, a suspicious email) sit at the top, scenarios that require a structural judgement about trust (USB provenance, Wi-Fi exposure model) sit lower, and the scenario in which all four options look surface-plausible (password) sits at the bottom. The data are consistent with the view that scenario surface, rather than topic difficulty alone, shapes in-the-moment decisions.

7.3 Design iteration as a finding

The pivot from the Ludii-based board-game prototype (Section 3.1) to the scenario-based web prototype that produced the present results is treated here as a design finding rather than a methodological footnote.

The first prototype framed each threat as a piece-and-board abstraction on a partial-information chess-like board, played against a Monte Carlo Tree Search baseline. Players acquired strategy under the abstraction without the strategy translating into the targeted security concepts: the chess-like setting rewarded inference and search over recognition of the specific threats the scenarios were meant to teach.

This aligns with a distinction established in the literature: strategic abstractions transfer best when the target competence is itself strategic (resource allocation, defence-in-depth planning, system-level threat modelling), whereas scenario-based designs anchored in plausible everyday moments transfer best when the target competence is recognition of specific cues in concrete situations [13], [19], [21], [27]. The five learning objectives sit firmly in the situated-recognition category, yet the Ludii prototype pursued them through a strategic-transfer medium. The playtest evidence surfaced that mismatch, and the re-implementation as a scenario-based web artefact is its consequence. The empirical results are therefore the second component of the contribution; the first is the design iteration that selected the medium in which those results could be produced at all.

7.4 Contributions in context

The work adds three things to the scenario-based and gamified-training literature. **Artefact and matched data infrastructure:** CyberNet ships as a single browser-deployed application that operates without installed software or classroom scaffolding, and logs scene-level decisions and confidence-marked test responses into the same event store as the test instruments; comparable phishing-awareness designs typically report knowledge change without an in-game decision signal [13], [15], [27]. **Calibration result:** most scenario-based cybersecurity games report accuracy change as the primary outcome [4], [14], [15]; the binary confidence marker on every item lets the present work show that confidently-wrong responses fall by roughly half rather than being relabelled as uncertain. **Presentation-layer probe:** the survey separates narration, music, and visual items from the TAM constructs, so the cinematic surface is evaluated as a distinct dimension. None of these claims that the prototype outperforms a specific comparable; the single-group design rules that out by construction.

7.5 Threats to validity

Internal validity remains the dominant constraint: the within-subject design cannot separate the intervention's contribution from a test-retest effect on the same ten items reused in the post-test. The no-feedback structure blocks the answer-memorisation form of the memory effect, since correct answers are never surfaced between the two tests; the residual concern is item familiarity. *Construct validity* is bounded by ten items across five concepts, and the per-concept comparison in particular is exploratory. *External validity* reflects convenience sampling through student communities and tech-leaning channels, which produced a younger sample (94 of 102 respondents in the 18–34 bands); the consistent gain across all five concepts suggests the prototype is not over-targeted to a single sub-population, but the sample cannot speak to populations outside the recruitment channels used. *Durability* is not measured: the single-session design captures the immediate post-test only.

8 Conclusion and Future Work

This paper set out to test whether a short, scenario-based web game could measurably improve everyday cybersecurity awareness, and the evidence says it can: across 102 participants, knowledge of all five concept areas rose significantly, the aggregate effect was large ($d = 1.07$), and the share of confidently-wrong answers fell by roughly half. Participants found the experience useful and engaging. The artefact is a single Next.js application that doubles as a public demonstration and as the study instrument, with runtime dependencies limited to the static client and an event database.

Four limitations bound the result. The single-group within-subject design cannot fully separate the intervention's effect from a test-retest effect on the reused items. The ten-item instrument samples two

items per concept, so the per-concept comparison is exploratory, and the absence of untaught control items prevents isolating concept-targeted learning from a generic test-retest effect. The single-session timeline is silent on retention. Convenience sampling produced a younger, tech-leaning sample. A within-session knowledge–behaviour gap also remains visible in the password scenario, where post-test accuracy of 73.0% sits well above the in-scenario safe-decision rate of 53.9%.

Five extensions follow. **A control-group replication** would pair the scenario-based intervention with a content-matched control (a short text article on the same five concepts) under between-subjects randomisation, and estimate test-retest effects through a no-intervention arm. **A delayed retention measurement** at two to four weeks would test durability, re-using the existing instrument with a re-consent step. **A broader knowledge instrument** with untaught control items would sharpen construct validity. **A two-stage instrument that probes behaviour after the debrief** would quantify how much the debrief shifts behaviour within the same session, by inserting a second decision scene in a slightly different surface before the post-test. **An LLM agent-persona replay** would seed personas from the demographic form, drive an automated agent through the same five scenarios, and compare its choices against the matched human participant’s actual choices in a matched-pairs design. The prototype, study instrument, and analysis pipeline are ready to support each of these.

Availability

The deployed web application is publicly available at <https://cybernet-awareness.vercel.app/>.

Acknowledgements

The author thanks everyone who took the time to play through the scenarios and complete the pre- and post-tests. This work would not exist without them.

Use of large language models

Large language models were used during the preparation of this work to refine wording, help organise the literature review, and scaffold sections of code that the author then reviewed and edited. They were not used to generate, alter, or interpret the study results: every number reported here comes from a deterministic analysis script run on the participant event log. The research questions, design choices, interpretations, and conclusions are the author’s own.

References

- [1] L. Hadlington, ‘Human factors in cybersecurity; examining the link between internet addiction, impulsivity, attitudes towards cybersecurity, and risky cybersecurity behaviours,’ *Heliyon*, vol. 3, no. 7, e00346, Jul. 2017. DOI: 10.1016/j.heliyon.2017.e00346
- [2] J. Dinét, *Human factors and information security: Impact of individual, social and culture environment*, Feb. 2016. DOI: 10.13140/RG.2.1.1172.4561
- [3] C. Y. Ng and M. K. B. Hasan, ‘Cybersecurity serious games development: A systematic review,’ *Computers & Security*, vol. 150, p. 104307, 2025. DOI: 10.1016/j.cose.2024.104307 [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167404824006138>
- [4] P. Bitrián, I. Buil, S. Catalán and D. Merli, ‘Gamification in workforce training: Improving employees’ self-efficacy and information security and data protection behaviours,’ *Journal of Business Research*, vol. 179, p. 114685, 2024. DOI: 10.1016/j.jbusres.2024.114685

- [5] R. Salameh and C. Loh, 'Engagement and players' intended behaviors in a cybersecurity serious game,' *International Journal of Gaming and Computer-Mediated Simulations*, vol. 14, no. 1, pp. 1–21, 2022. DOI: 10.4018/IJGCMS.313185
- [6] C. Moumouh, J. García Berná, M. Y. Chkouri and J. Fernández-Alemán, 'Serious games to improve privacy and security knowledge for professionals: A systematic literature review,' *International Journal of Serious Games*, vol. 12, no. 1, pp. 3–24, Feb. 2025. DOI: 10.17083/ijsg.v12i1.825
- [7] A. Gwenhure and F. Rahayu, 'Gamification of cybersecurity awareness for non-it professionals: A systematic literature review,' *International Journal of Serious Games*, vol. 11, no. 1, pp. 83–99, Mar. 2024. DOI: 10.17083/ijsg.v11i1.719
- [8] A. Alnajim, S. Habib, M. Islam, H. Alrawashdeh and M. Wasim, 'Exploring cybersecurity education and training techniques: A comprehensive review of traditional, virtual reality, and augmented reality approaches,' *Symmetry*, vol. 15, no. 12, p. 2175, 2023. DOI: 10.3390/sym15122175
- [9] O. Pahlavanpour and S. Gao, 'A systematic mapping study on gamification within information security awareness programs,' *Heliyon*, vol. 10, no. 19, e38474, 2024. DOI: 10.1016/j.heliyon.2024.e38474
- [10] F. D. Davis, 'Perceived usefulness, perceived ease of use, and user acceptance of information technology,' *MIS Quarterly*, vol. 13, no. 3, pp. 319–340, 1989. DOI: 10.2307/249008
- [11] R. Garris, R. Ahlers and J. E. Driskell, 'Games, motivation, and learning: A research and practice model,' *Simulation & Gaming*, vol. 33, no. 4, pp. 441–467, 2002. DOI: 10.1177/1046878102238607 eprint: <https://doi.org/10.1177/1046878102238607>. [Online]. Available: <https://doi.org/10.1177/1046878102238607>
- [12] J. Gee, 'Video games, mind, and learning,' *The Interactive Digital Media & Arts Association Journal*, vol. 2, Jan. 2005. [Online]. Available: https://www.researchgate.net/publication/228362406_Video_games_mind_and_learning
- [13] N. Henderson, H. Pallett, S. van der Linden, J. Montanarini and O. Buckley, 'The disphishinformation game: Creating a serious game to fight phishing using blended design approaches,' Jan. 2024. DOI: 10.54941/ahfe1004774
- [14] T. Steen and J. Deeleman, 'Successful gamification of cybersecurity training,' *Cyberpsychology, Behavior, and Social Networking*, vol. 24, no. 9, pp. 593–598, Sep. 2021. DOI: 10.1089/cyber.2020.0526
- [15] E. Le-Nye, C. Yaacoub and J. Possik, 'Evaluating phishing awareness strategies: A comparative study of education-based approaches and game-based learning,' *Procedia Computer Science*, vol. 251, pp. 666–671, Jan. 2024. DOI: 10.1016/j.procs.2024.11.166
- [16] R. Hodhod, H. Hardage, S. Abbas and E. Aldakheel, 'Cyberhero: An adaptive serious game to promote cybersecurity awareness,' *Electronics*, vol. 12, no. 17, p. 3544, 2023. DOI: 10.3390/electronics12173544
- [17] G. P. Kusuma, E. K. Wigati, Y. Utomo and L. K. Putera Suryapranata, 'Analysis of gamification models in education using mda framework,' *Procedia Computer Science*, vol. 135, pp. 385–392, 2018, The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life, ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2018.08.187> [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050918314765>
- [18] D. V and A. Yadav, 'Constructivism: A paradigm for teaching and learning,' *Arts and Social Sciences Journal*, vol. 7, Jan. 2016. [Online]. Available: https://www.researchgate.net/publication/308084623_Constructivism_A_Paradigm_for_Teaching_and_Learning
- [19] S. Hart, B. Halak and V. Sassone, 'Motens: A pedagogical design model for serious cyber games,' Oct. 2021. [Online]. Available: https://www.researchgate.net/publication/365568874_MOTENS_a_pedagogical_design_model_for_serious_cyber_games

- [20] H. Taherdoost, 'Towards an innovative model for cybersecurity awareness training,' *Information*, vol. 15, no. 9, p. 512, 2024. DOI: 10.3390/info15090512
- [21] S. Arnab et al., 'Mapping learning and game mechanics for serious games analysis,' *British Journal of Educational Technology*, vol. 46, no. 2, pp. 391–411, 2015. DOI: 10.1111/bjet.12113
- [22] K. Kronhardt, K. Rolfes and J. Gerken, 'Trickery: Exploring a serious game approach to raise awareness of deceptive patterns,' in *Proceedings of the 23rd International Conference on Mobile and Ubiquitous Multimedia (MUM '24)*, New York, NY, USA: Association for Computing Machinery, 2024, pp. 133–147, ISBN: 9798400712838. DOI: 10.1145/3701571.3701588
- [23] M. Katsantonis, A. Manikas, N. Partarakis and I. Mavridis, 'Yellow training: A collaborative tocofelet tabletop serious game for cybersecurity awareness and training,' *International Journal of Information Security*, vol. 24, no. 4, p. 171, 2025. DOI: 10.1007/s10207-025-01086-0
- [24] K. Labunets, F. Massacci, F. Paci and L. M. S. Tran, 'An experimental comparison of two risk-based security methods,' in *2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, Oct. 2013, pp. 163–172. DOI: 10.1109/ESEM.2013.29
- [25] T. Leng, *Scenario selection and student assessment modules for cyberciege*, Dec. 2003. [Online]. Available: https://www.researchgate.net/publication/235061879_Scenario_Selection_and_Student_Assessment_Modules_for_CyberCIEGE
- [26] C. Irvine, M. F. Thompson and K. Allen, 'CyberCIEGE: Gaming for information assurance,' *IEEE Security & Privacy*, vol. 3, no. 3, pp. 61–64, 2005. DOI: 10.1109/MSP.2005.64
- [27] L. Kassner and A. Schönbohm, 'A serious game to improve phishing awareness,' in *Games and Learning Alliance*, Cham: Springer International Publishing, 2022, pp. 109–117. DOI: 10.1007/978-3-031-22124-8_11
- [28] K. Beckers and S. Pape, 'A serious game for eliciting social engineering security requirements,' in *2016 IEEE 24th International Requirements Engineering Conference (RE)*, 2016, pp. 16–25. DOI: 10.1109/RE.2016.39
- [29] A. Yasin, L. Liu, T. Li, R. Fatima and J. Wang, 'Improving software security awareness using a serious game,' *IET Software*, vol. 13, no. 2, pp. 159–169, 2019. DOI: 10.1049/iet-sen.2018.5095
- [30] T. Denning, A. Lerner, A. Shostack and T. Kohno, 'Control-alt-hack: The design and evaluation of a card game for computer security awareness and education,' in *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2013, pp. 915–928. DOI: 10.1145/2508859.2516753
- [31] T. Denning, A. Shostack and T. Kohno, 'Practical lessons from creating the Control-Alt-Hack card game and research challenges for games in education and research,' in *USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE)*, San Diego, CA, 2014. [Online]. Available: <https://www.usenix.org/conference/3gse14/summit-program/presentation/denning>
- [32] A. Dabrowski, M. Kammerstetter, E. Thamm, E. Weippl and W. Kastner, 'Leveraging competitive gamification for sustainable fun and profit in security education,' in *USENIX Summit on Gaming, Games, and Gamification in Security Education (3GSE)*, Washington, D.C., 2015. [Online]. Available: <https://www.usenix.org/conference/3gse15/summit-program/presentation/dabrowski>
- [33] E. Trickel et al., 'Shell we play a game? ctf-as-a-service for security education,' Aug. 2017. [Online]. Available: https://www.researchgate.net/publication/319141725_Shell_We_Play_A_Game_CTF-as-a-service_for_Security_Education

- [34] S. Frey, A. Rashid, P. Anthony, M. Albuquerque and S. A. A. Naqvi, ‘The good, the bad and the ugly: A study of security decisions in a cyber-physical systems game,’ in *Proceedings of the 40th International Conference on Software Engineering (ICSE ’18)*, May 2018, p. 496, ISBN: 978-1-4503-5638-1. DOI: 10.1145/3180155.3182549
- [35] A. Spatafora, M. Wagemann, C. Sandoval, M. Leisenberg and C. Carvalho, ‘An educational escape room game to develop cybersecurity skills,’ *Computers*, vol. 13, no. 8, p. 205, 2024. DOI: 10.3390/computers13080205
- [36] M. Wa Nkongolo, ‘Infusing morabaraba game design to develop a cybersecurity awareness game (cybermoraba),’ in *International Conference on Cyber Warfare and Security*, vol. 19, 2024, pp. 240–250. DOI: 10.34190/iccws.19.1.1957
- [37] D. N. Chandrashekar, A. Lee, M. Azab and D. Gracanin, ‘Understanding user behavior for enhancing cybersecurity training with immersive gamified platforms,’ *Information*, vol. 15, no. 12, p. 814, 2024. DOI: 10.3390/info15120814
- [38] G. Tempestini, S. Merà, M. P. Palange, A. Bucciarelli and F. Di Nocera, ‘Improving the cybersecurity awareness of young adults through a game-based informal learning strategy,’ *Information*, vol. 15, no. 10, p. 607, 2024. DOI: 10.3390/info15100607
- [39] É. Piette, D. J. N. J. Soemers, M. Stephenson, C. F. Sironi, M. H. M. Winands and C. Browne, ‘Ludii – the ludemic general game system,’ in *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, 2020, pp. 411–418. DOI: 10.3233/FAIA200120
- [40] M. Tischer et al., ‘Users really do plug in USB drives they find,’ in *2016 IEEE Symposium on Security and Privacy (SP)*, IEEE, 2016, pp. 306–319. DOI: 10.1109/SP.2016.26
- [41] N. Cheng, X. O. Wang, W. Cheng, P. Mohapatra and A. Seneviratne, ‘Characterizing privacy leakage of public WiFi networks for users on travel,’ in *2013 Proceedings IEEE INFOCOM*, IEEE, 2013, pp. 2769–2777. DOI: 10.1109/INFOCOM.2013.6567086
- [42] J. Fu, Y. Lu, Z. Yang, F. Nah and R. LC, ‘Cracking Aegis: An adversarial LLM-based game for raising awareness of vulnerabilities in privacy protection,’ in *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS ’25)*, 2025, pp. 639–662. DOI: 10.1145/3715336.3735812 arXiv: 2505.16954 [cs.HC].