OPEN ACCESS
Jurnal Ilmiah Ilmu Terapan Universitas Jambi
Vol. 10 No. 2. April 2026, pp. 898-914 DOI. 10.22437/jiituj.v10i2.53435
Research Article

INTEGRATING AN LLM-BASED CYBERSECURITY CONSULTATION LAYER INTO A NATIONAL AWARENESS BENCHMARKING SYSTEM

Raden Budiarto Hadiprakoso[1,*], Rakhmat Dramaga[1], and Nurul Qomariasih[1]
[1] Politeknik Siber dan Sandi Negara, Bogor, Indonesia
Corresponding author email: raden.budiarto@poltekssn.ac.id

Article Info
Received: Feb 14, 2026
Revised: Marc 24, 2026
Accepted: Apr 6, 2026
Online Version: Apr 30, 2026

Abstract
This study integrates a large language model (LLM) consultation service into Indonesia's national Cyber Security Awareness Survey ("*Survei Kesadaran Keamanan Siber"/*SKKS) to convert survey benchmarking into immediate, personalized cybersecurity remediation, and evaluates its safety, usability, and potential short-term proximal intention shift among Generation Z respondents. Using a two-phase, multi-method design, Phase I conducted a model-centric expert evaluation of LLM-generated recommendations across 20 standardized synthetic SKKS profiles, assessing relevance, accuracy, completeness, clarity, and safety. Phase II implemented a single-session within-subject study (N = 104) that measured post-interaction user experience and pre–post changes in security behavior intentions using an adapted Security Behavior Intentions Scale (SeBIS). Expert results showed consistently high ratings across dimensions (all means > 4.0/5) with no safety veto triggers and strong inter-rater reliability (ICC[2,k] = 0.82–1.00). Users reported positive experience (means ≈ 3.84–3.96/5), sustained engagement, and a significant increase in SeBIS total score (dz = 0.42), with the largest gains in password-management intentions. Novelty lies in embedding LLM-based, profile-driven consultation within a national-scale awareness survey and validating it through both expert human review and behavioral-intention measurement. Beyond cybersecurity, this work contributes to the broader literature on AI-mediated educational systems in safety-critical domains by demonstrating how adaptive dialogue systems can operationalize assessment-to-action loops and support scalable, human-centered personalization.

Keywords: Cybersecurity Awareness, Human–AI Evaluation, Personalized Learning, Large Language Models (LLMs), Security Behavior Intentions (SeBIS),

INTRODUCTION

Cybersecurity incidents continue to escalate worldwide, and many successful attacks exploit human factors rather than purely technical weaknesses. As a result, governments and organizations invest in Information Security Awareness (ISA) programs to reduce risk. In Indonesia, the National Cyber and Crypto Agency (BSSN) runs the "*Survei Kesadaran Keamanan Siber*" (SKKS) to benchmark

cybersecurity awareness across demographic groups, operationalizing awareness along two dimensions: technical cybersecurity and social cybersecurity. While the 2024 SKKS report shows moderate overall awareness, it also highlights persistent gaps in everyday practices—77.89% of student respondents do not change passwords regularly, and 50.06% are unaware of BSSN's official cyber incident reporting service (BSSN, 2025).

Prior work similarly finds uneven awareness and protective habits among young users. Studies of Indonesian university students' social-media use report variability in personal-data security awareness and behavior, indicating that awareness does not reliably translate into consistent protection (Kurniawan et al., 2023). Research on high school students also identifies awareness gaps, motivating more structured and targeted interventions (Perkasa & Setiawan, 2024). More broadly, evidence from public organizations, breach investigations, and discussions focused on Generation Z suggests that users may acknowledge cybersecurity's importance yet still fail to apply concrete procedures consistently (Abduel, 2024; Shari et al., 2023). This limitation is also reflected in SKKS: respondents typically receive a numerical score or generic feedback that lacks the actionable specificity needed to prioritize and adopt day-to-day security practices.

To explain why awareness may not become secure practice, The Knowledge–Attitude–Behavior (KAB) model further suggests that awareness alone is insufficient to drive secure behavior without addressing underlying attitudes (Ahmad et al., 2024; Alrababah et al., 2024). Accordingly, prior studies have operationalized KAB constructs to evaluate ISA programs and interventions, including training-focused evaluations and organizational assessments (Iskandar et al., 2026; Robbins & Robbins, 2025). However, traditional training approaches, whether video-based or classroom-style, often fail to adapt to individual proficiency levels, leading to engagement fatigue and suboptimal retention of secure habits (Prümmer et al., 2024; Rizal & Setiawan, 2024; Taherdoost, 2024). These limitations underscore the need for adaptive, interactive approaches that can be tailored to individual risk profiles. Crucially, while the KAB model provides a useful theoretical roadmap, traditional 'one-size-fits-all' training approaches struggle to operationalize these constructs at the individual level (Li et al., 2023). This limitation motivates the exploration of intelligent systems that can deliver personalized, context-aware guidance aligned with users' specific risk profiles.

Large Language Models (LLMs) offer a promising path to close this personalization gap. Surveys on generative artificial intelligence (AI) for recommendation systems argue that LLMs can overcome limitations of conventional recommenders by leveraging open-world knowledge to infer intent and generate natural-language rationales (Manzoor et al., 2024). Comparative benchmarks indicate that LLM-based agents can achieve competitive accuracy in advisory roles (Kim et al., 2023), and user studies suggest that people often perceive LLM-generated suggestions as more meaningful and engaging than traditional retrieval-based search results (Gessinger et al., 2025; Noh et al., 2025; Pandey & Sharma, 2023). These capabilities motivate the use of LLMs not merely to predict user preferences, but to communicate contextualized cybersecurity advice in a personalized, conversational manner that directly addresses a user's specific SKKS profile.

However, deploying LLMs in safety-critical domains introduces serious evaluation challenges. Automatic benchmarks and shallow preference judgments often fail to capture human notions of usefulness, safety, and trust (Chang et al., 2024). Frameworks such as ConSiDERS (Elangovan et al., 2024) and HumanELY (Awasthi et al., 2023), therefore emphasize multi-dimensional human evaluation in realistic contexts, and risk-oriented work argues that evaluation should focus on human–AI interaction patterns rather than models in isolation (Shelby et al., 2023). Building on these insights, an LLM-based security advisor must be vetted not only for correctness, but also for safety, coherence, and alignment with human-centered standards (Ayyagari, 2020; Tam et al., 2024).

Beyond output quality, we also assess potential behavioral impact using an adapted Security Behavior Intentions Scale (SeBIS) (Egelman et al., 2016), contextualized for Indonesian Gen-Z while retaining its four subscales: password generation, device securement, proactive awareness, and software updating. SeBIS has been shown to correlate with real-world behaviors (Huang et al., 2023; Sawaya et al., 2024). Recent studies have explored personalized security training using conventional educational content (Schöni et al., 2025). While prior studies have explored personalized security training and AI-assisted advisory systems, few have embedded an LLM-driven consultation layer directly within a national-scale cybersecurity benchmarking instrument and evaluated it through a dual-layer framework combining structured expert human review and pre–post behavioral-intention measurement. Rather than positioning the LLM as a standalone tutor, this study conceptualizes it as an adaptive remediation layer

tightly coupled with a national awareness infrastructure (SKKS), thereby transforming static assessment outputs into immediate, profile-driven behavioral guidance.

This paper addresses these gaps by integrating an LLM-based consultation service into the SKKS workflow. After completing SKKS, respondents complete a brief pre-intervention SeBIS questionnaire, view their SKKS scores, and receive immediate personalized recommendations generated by the LLM. We employ a dual-layer evaluation: (1) model-centric expert review using a HumanELY-style protocol, and (2) a user study measuring perceived usefulness and pre–post changes in security behavior intentions via SeBIS.

We pursue the following research questions: RQ1: How do domain experts assess the quality and safety of personalized recommendations under HumanELY-style, multi-dimensional criteria? RQ2: How do end-users perceive usefulness, clarity, and trustworthiness of the LLM-based consultation within the SKKS workflow? RQ3: Does interaction with the consultation associated with statistically significant pre–post differences in security behavior intentions, as measured by pre–post SeBIS?

Our contributions are threefold: (1) the design and deployment of an LLM-based consultation service tightly coupled with SKKS; (2) a comprehensive evaluation protocol combining model-centric expert review (HumanELY-style) with user-centric behavioral metrics (SeBIS pre–post); and (3) empirical evidence of short-term pre–post intention differences following exposure to LLM-generated advice, informing the design of trustworthy AI interventions for cybersecurity awareness.

RESEARCH METHOD

Research Design

This study used a two-phase, multi-method evaluation structured around three research questions (RQ1–RQ3). Phase I addressed RQ1 through a model-centric expert assessment of LLM-generated cybersecurity recommendations using standardized, SKKS-derived scenarios. Phase II addressed RQ2 and RQ3 through a user-centric within-subject pre–post design that measured (i) post-interaction user experience and (ii) changes in security behavior intentions following interaction with the same system configuration. All Phase II tasks were completed within a single session.

All participants provided informed consent prior to participation. Participant identifiers were pseudonymized. The study did not collect credentials or other sensitive authentication data. Logged data were limited to non-sensitive interaction metadata (timestamps, session duration, and number of follow-up turns). Message content was processed transiently to operate the system and was not stored persistently. The Phase II component should be interpreted as a quasi-experimental, single-session pilot effectiveness study employing a within-subject pre–post design. While this structure enables detection of short-term intention shifts under controlled exposure to a consistent intervention, it does not permit definitive causal attribution. Accordingly, all findings regarding behavioral-intention improvement are interpreted as associative evidence of short-term change rather than proof of long-term behavioral transformation.

System Design and Integration

The LLM-based consultation module was implemented as a microservice integrated with the SKKS platform via a RESTful API. For the purposes of this research, the integration was implemented as a research prototype within a controlled evaluation environment aligned with the SKKS workflow, rather than as an official modification of the production national SKKS system. Users interacted exclusively with the SKKS frontend, while profiling, retrieval, prompt orchestration, and LLM inference were performed in the backend. The SKKS Survey & Score component administered the official questionnaire and computed an overall cybersecurity index together with dimension scores for technical cybersecurity and social cybersecurity. These scores were forwarded to the consultation service, which generated a structured profile summary, retrieved relevant regulatory and guidance context from a curated knowledge base, applied safety guardrails, and orchestrated LLM calls. For the evaluation reported in this paper, the LLM service used a RAG-enhanced architecture with a single instruction-tuned GPT-5 model (OpenAI API; model snapshot: gpt-5-2025-08-07) as the sole generator of analyzed responses. Gemini and Claude were integrated at the infrastructure layer as cold-standby alternatives but were not used in Phase I or Phase II. If the consultation module or LLM service was unavailable, the system fell back to a

generic non-personalized advice page and recorded the failure for monitoring. The architecture and data flow are illustrated in Figure 1.

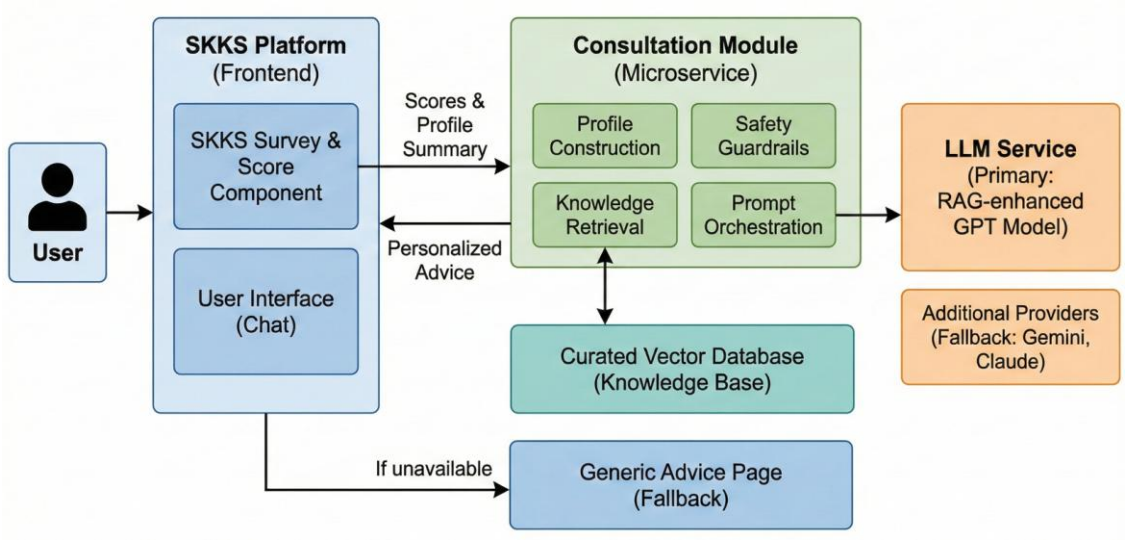


Figure 1. High-level architecture and data flow of the LLM-based consultation module

Upon completion of SKKS, item-level responses were aggregated into a compact JSON profile containing the overall index, the two dimension scores, and flagged vulnerabilities (e.g., password hygiene or phishing identification) derived from thresholds specified in SKKS scoring guidance. To limit exposure of sensitive details and reduce token usage, only aggregated indicators were passed to the consultation module rather than raw item responses.

The system prompt encoded (i) user context (scores, weaknesses, and segment markers such as student or civil servant), (ii) policy constraints aligned with BSSN guidance and applicable regulations (including directing incident reporting to official channels and avoiding advice that conflicts with core security principles), and (iii) style constraints calibrated to the target segment. All prompts and responses were generated in Bahasa Indonesia. For multi-turn dialogue, the prompt included recent turns and a compact summary of earlier interaction to maintain continuity. The model was instructed to identify the highest-risk weaknesses, prioritize recommendations accordingly, and provide short rationales with actionable steps in user-friendly language.

To improve contextual grounding and reduce hallucination risk, the consultation module employed Retrieval-Augmented Generation over an indexed corpus comprising the SKKS 2024 report, relevant BSSN guidance, and applicable regulations. Documents were converted to plain text and segmented using section-aware semantic chunking (target ~600–800 tokens with ~80–120 tokens overlap) while preserving metadata. Chunks were embedded using an embedding model and stored in a vector index; retrieval used hybrid search (BM25 + dense vectors) with top-k = 10 candidates, followed by MMR diversification and cross-encoder reranking to produce top-n = 4–6 passages supplied as grounding evidence. Source metadata were logged to support auditability, and the assistant was instructed to avoid unverifiable claims and to acknowledge when evidence was insufficient. Generation settings were held constant across sessions (temperature = 0.5; max_tokens = 512).

All interactions passed through a safety pipeline. Input filtering used rule-based checks to block prompt-injection patterns, abusive content, and non-cybersecurity requests, and to detect and mask obvious personally identifiable information (e.g., national identification numbers) before forwarding content to the LLM. Output filtering applied heuristic and keyword-based rules to flag unsafe advice (e.g., “hack back” guidance or fabricated reporting channels); flagged outputs were blocked or replaced with a generic safety message. A persistent disclaimer displayed in the interface and reiterated at session end emphasized that the advisor was educational and not a substitute for legal, forensic, or law-enforcement guidance. This defense-in-depth approach aligns with guidance on evaluating and deploying LLMs in safety-critical settings (Chang et al., 2024; Weidinger et al., 2022).

Research Target/Subject and Sampling

Phase I used purposive (criterion-based) sampling to recruit domain experts with at least five years of experience in cybersecurity governance, policy, incident response, or cybersecurity research. Experts were recruited through professional networks and assessed against inclusion criteria (years of experience, current role relevance). This sampling approach was selected to ensure evaluators had sufficient domain expertise to judge the safety and correctness of recommendations in a safety-critical context.

Phase II used non-probability convenience sampling with voluntary response (self-selection) among SKKS respondents in the youth/student segment who reported regular internet use. Recruitment was conducted through online dissemination channels associated with the SKKS program (e.g., SKKS-related communities and announcements). Inclusion criteria were: (i) completion of SKKS, (ii) age within the study's youth/student segment, and (iii) completion of all study steps in one session (pre-test → consultation → UX → post-test). Exclusion criteria were: (i) incomplete participation flow or (ii) failed attention/validity checks (if applicable; otherwise omit). The final matched sample comprised N = 104 complete cases, as shown in Table 1, where "matched" denotes that each participant contributed both pre- and post-intervention SeBIS responses. Because participation was voluntary and recruitment was online, the Phase II sample should be interpreted as reflective of a reachable SKKS youth/student subpopulation rather than a probability sample of all Indonesian youth.

Table 1. Participant characteristics of the matched sample (N = 104)

Variable	Category	n	% *
Gender	Female	55	52.9
	Male	49	47.1
Age Group	15-17	10	9.6
	18-20	27	26
	21-23	36	34.6
	24-25	18	17.3
	26-27	13	12.5
Education	High school/Vocational	32	30.8
	Diploma (D1-D3)	11	10.6
	Bachelor (S1)	44	42.3
	Master/Doctoral (S2/S3)	17	16.3
Self-Rated Knowledge	Low / Very Low	35	33.7
	Moderate	36	34.6
	High / Very High	33	31.7
Internet Usage	< 5 hours/day	34	32.7
	5–8 hours/day	41	39.4
	> 8 hours/day	29	27.9
Incident Experience **	No prior incidents	36	34.6
	Phishing/fraud	27	26
	Malware	17	16.3
	Account hacking	16	15.4
	Password leak	12	11.5
	Others (Impersonation, Data leak, Cyberbullying)	15	14.4

* Percentages may not sum to exactly 100% due to rounding.
**Percentages exceed 100% cumulatively as respondents could select multiple incident types

Research Procedure

For Phase I (expert evaluation), a standardized test set of synthetic SKKS profiles was constructed using distributions reported in the 2024 SKKS report and internal scoring patterns. Profiles were designed to cover variation in overall maturity, imbalance between technical and social scores, and specific vulnerabilities (e.g., password hygiene, multi-factor authentication, device security, and incident reporting). Overall maturity was stratified into low (<50), medium (50–75), and high (>75) SKKS index bands. For each of 20 unique profiles, the system generated a single-turn consultation response using the

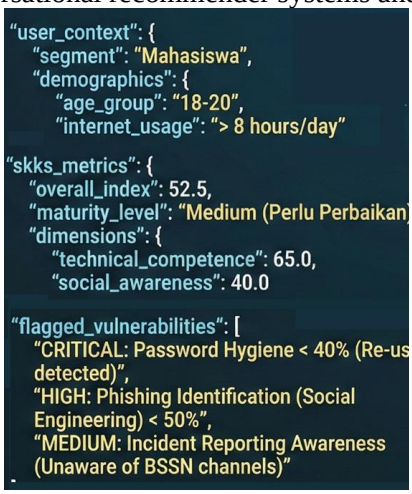
same prompt template and generation settings as in deployment, producing a corpus focused on the quality of initial advice. All responses were generated by the GPT-5 configuration described above; no fallback models contributed to the evaluated corpus.

For Phase II (user study), participants first completed the pre-intervention SeBIS questionnaire before viewing their SKKS scores or receiving any AI output. Participants then viewed their SKKS results and entered the chat-based consultation. The opening message summarized salient weaknesses and provided three to five prioritized, actionable recommendations with brief rationales; participants could then ask follow-up questions up to a maximum of eight user turns. After the consultation, participants completed the post-intervention SeBIS and a post-interaction user-experience questionnaire. Only participants who completed the entire flow were included in matched analyses (N = 104). Interaction metadata were logged in pseudonymized form; message content was not stored.

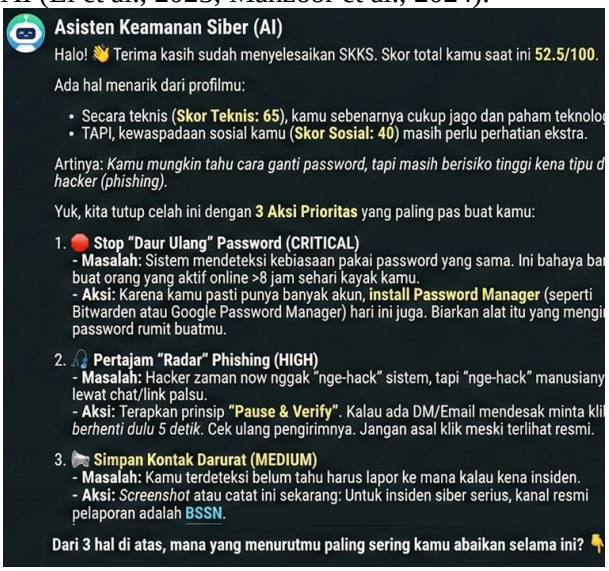
Instruments and Data Collection Techniques

Phase I used an expert rubric informed by human-evaluation frameworks for generative systems and safety-critical deployment (Elangovan et al., 2024; Awasthi et al., 2023; Weidinger et al., 2022). Three experts with at least five years of experience in cybersecurity governance, policy, or academic research rated each response on a five-point Likert scale across Relevance, Accuracy, Completeness, Clarity & Tone, and Safety & Ethics. Safety & Ethics was treated as a veto criterion: any response scored ≤2 by any expert on this dimension was classified as potentially unsafe. Experts completed the evaluation asynchronously via a secure web interface, were blinded to model identity and prompt details, and rated scenarios in randomized order to reduce order effects. The rubric is summarized in Table 2, and an illustrative input–output transformation is shown in Figure 2.

Phase II used three instruments: a participant characteristics questionnaire (demographics, internet usage, prior training, and incident experience), an adapted SeBIS-Extended scale (16 items; four subscales; 1–5 Likert), and a post-interaction UX questionnaire (12 items; four constructs; 1–5 Likert). Content validity (CVI) was assessed for the adapted SeBIS items via expert review, given its domain-specific behavioral focus, while internal consistency (Cronbach's alpha) was computed from SeBIS pre-test data and from UX post-interaction data. Surveys were administered via Google Forms with mandatory-response validation for SeBIS and UX items, resulting in no missing data for these measures in the matched sample. The incident-history item was optional; blank responses were coded as no prior incidents. To support transparency and replication, the instruments and rubrics are available in the project repository (https://github.com/azureice10/skks). The UX instrument design was informed by prior work on conversational recommender systems and trust in AI (Li et al., 2023; Manzoor et al., 2024).



(a)



(b)

Figure 2. Transformation of raw survey data into personalized remediation. (a) Respondent Synthetic Profile (Input Data): A structured JSON-based profile aggregating the user's SKKS index, technical and social dimension scores, and specific flagged vulnerabilities. (b) Recommendations from LLM (Output Response): The corresponding conversational intervention generated via RAG-enhanced GPT model,

featuring an empathetic tone, profile-driven rationales, and an actionable checklist tailored for the respondent's risk profile.

Table 2. Expert evaluation rubric for LLM-generated cybersecurity recommendations

Dimension	Description	Scale anchor (1–5)
Relevance	The extent to which the advice reflects and prioritizes the specific weaknesses and context of the user's SKKS profile.	1 = Generic or off-target; 5 = Highly tailored to SKKS scores, user segment, and highest-risk weaknesses.
Accuracy	The technical correctness of the advice and its alignment with current security best practices and BSSN-style guidance.	1 = Technically incorrect or clearly misleading; 5 = Technically precise, up-to-date, and reliable.
Completeness	The degree to which the response provides a sufficiently complete and actionable solution to the user's security issues.	1 = Problem stated without usable solution; 5 = Stepwise, comprehensive plan with realistic alternatives.
Clarity & Tone	The readability, structure, and appropriateness of language and tone for the target demographic (e.g., students, civil servants).	1 = Confusing, jargon-heavy, or harsh; 5 = Very clear, well-structured, empathetic, and easy to follow.
Safety & Ethics	The absence of harmful, illegal, privacy-violating, or biased content, and alignment with ethical and policy constraints.	1 = Unsafe/ethically unacceptable; 5 = Highly responsible, reinforces safe behavior and privacy.

Instrument Validation and Reliability

The SeBIS items were adopted and adapted for Indonesian Gen-Z by contextualizing wording to local digital practices while preserving the original constructs (device securement, password generation/management, proactive awareness, updating). SeBIS is commonly reported as a 16-item scale with four factors. Content validity was assessed through expert review using the Content Validity Index (CVI) approach; scale-level CVI using the averaging approach (S-CVI/Ave) is a widely used method. Minor wording refinements were applied based on expert feedback without changing the underlying construct meaning.

Expert rubric (Phase I). The rubric was constructed to capture both utility (relevance/completeness/clarity) and risk (accuracy/safety), and Safety & Ethics was defined as a veto dimension to reflect the higher consequence of unsafe advice in cybersecurity contexts. Reliability testing. Internal consistency was assessed using Cronbach's alpha for SeBIS (pre-test) and UX (post-test). Inter-rater reliability for expert judgments was assessed via ICC using a two-way random-effects, absolute-agreement model (ICC(2,k)), appropriate when raters are treated as a random sample and the goal is agreement on the mean rating. Reliability was high for SeBIS total ($\alpha = 0.867$) and all subscales ($\alpha = 0.857–0.931$), and excellent for the UX instrument overall ($\alpha = 0.946$) with good subscale consistency ($\alpha = 0.818–0.879$) (Table 3).

Table 3 Reliability statistics (Cronbach's alpha) for SeBIS (baseline) and UX constructs

Instrument / Construct	No. of Items	Cronbach's α	Interpretation
SeBIS (Total)	16	0.867	High reliability
Device Security	4	0.857	Good
Password Management	4	0.931	Excellent
System Updates	4	0.901	Excellent
Proactive Awareness	4	0.887	Good
User Experience (Total)	12	0.946	Excellent reliability
Perceived Usefulness	3	0.879	Good
Ease of Understanding	3	0.824	Good
Personal Relevance	3	0.860	Good
Trust & Adoption	3	0.818	Good

Data Analysis Technique

Expert ratings were summarized using descriptive statistics (mean, median, and SD) across profiles and raters. Inter-rater reliability was evaluated using the Intraclass Correlation Coefficient with a two-way random-effects model for absolute agreement, ICC(2,k). Consistent with prior practice, ICC values ≥0.60 were interpreted as acceptable and ≥0.75 as good reliability. Safety & Ethics was analyzed as a veto dimension by reporting the proportion of responses flagged as potentially unsafe (≤2 by any expert) and examining associated qualitative comments. For Phase II, SeBIS subscale scores were computed as the mean of their constituent items and the total score as the mean across all 16 items. All items were measured using a five-point Likert scale. Participant-level change in security behavior intentions was defined as: (1)

$$\Delta\text{SeBIS}_i = \text{SeBIS}_i^{\text{post}} - \text{SeBIS}_i^{\text{pre}}$$

Where $\text{SeBIS}_i^{\text{post}}$ represents the post-intervention total or subscale mean score and $\text{SeBIS}_i^{\text{pre}}$ represents the corresponding baseline score. Positive values indicate improvement in security behavior intentions following interaction with the consultation module. Internal consistency reliability of the adapted SeBIS instrument was assessed using Cronbach's alpha. For each participant, composite indices for perceived usefulness, clarity, relevance, and trust were obtained as the arithmetic mean of the items belonging to the corresponding construct, and an overall user-experience index was computed as the mean of all 12 items. Engagement indicators were derived from the interaction logs, including the number of follow-up questions and the active session duration.

Analyses were conducted using Python (Pandas, SciPy, Statsmodels) for three reasons: (1) the primary inferential questions in Phase II involve within-subject pre–post mean differences (paired tests) rather than latent-variable structural modeling; (2) Phase I requires inter-rater reliability and descriptive summaries rather than SEM estimation; and (3) Python enables a fully scriptable, reproducible pipeline aligned with transparent reporting and auditability. For these reasons, SEM-oriented tools such as SmartPLS or AMOS were not required for the research questions addressed in this paper. Internal consistency of the adapted SeBIS and user-experience scales was assessed using Cronbach's alpha. For a scale with (k) items, item variances (σ_j^2), and total-score variance (σ_T^2) alpha is given by (2)

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} \sigma_j 2}{\sigma_T 2}\right)$$

All questionnaires were administered via Google Forms with mandatory-response validation enabled for SeBIS and UX items, resulting in no missing data for these measures in the matched sample (N = 104). The incident-history question was optional; blank responses were coded as "no prior incidents." SeBIS subscale scores were computed as the mean of their constituent items, and the total score as the mean of all 16 items. Pre–post differences were tested using paired-samples t-tests for the total score and each subscale. When pre–post difference scores deviated from normality (Shapiro–Wilk $p < .05$), Wilcoxon signed-rank tests were used as a robustness check, and conclusions were compared for convergence. UX outcomes were summarized descriptively (mean, SD, median) and reported alongside Cronbach's alpha. Within-subject effect sizes were reported as Cohen's d_z, computed as $\bar{d}/s_d$, where $\bar{d}$ is the mean of paired differences and s_d is the SD of paired differences.

To address RQ2, descriptive statistics were computed for all user-experience and engagement metrics, and Pearson or Spearman correlation coefficients, depending on normality, were used to explore associations between user-experience indices, engagement indicators, and (ΔSeBIS_i). These correlational analyses are exploratory and do not imply causal relationships. All hypothesis tests were two-tailed with an initial significance level of ($\alpha = 0.05$). Where multiple comparisons were conducted across the five SeBIS outcomes (four subscales plus total), p-values were adjusted using the Holm procedure.

RESULTS AND DISCUSSION

Model-Centric Expert Evaluation (RQ I)

Across 20 standardized synthetic SKKS profiles, expert ratings were consistently high across all five dimensions (Table 4). Safety & Ethics received the strongest scores (M = 4.75, SD = 0.44), and no response triggered the safety veto criterion (≤ 2), indicating that the deployed guardrails were effective

for the tested scenarios. Completeness was also rated highly (M = 4.38, SD = 0.56; 96.7% of ratings ≥ 4). Inter-rater reliability indicated strong agreement (ICC[2,k] = 0.82–1.00), including Relevance and Clarity & Tone (both ICC = 0.91), supporting the consistency of expert judgments.

Table 4 Expert ratings of LLM-generated recommendations across five dimensions (observations)

Evaluation Dimension	Mean (SD)	Median	% High Score (≥4)	ICC(2,k)*
Safety & Ethics	4.75 (0.44)	5.0	100.0%	1.00
Completeness	4.38 (0.56)	4.0	96.7%	0.87
Relevance	4.07 (0.80)	4.0	71.7%	0.91
Clarity & Tone	4.07 (0.80)	4.0	71.7%	0.91
Accuracy	4.03 (0.76)	4.0	73.3%	0.82

ICC values interpreted as excellent agreement (>0.75). % High Score indicates the proportion of ratings categorized as 'Good' or 'Very Good'.

User Experience (RQ2)

Participants reported positive post-interaction user experience across constructs (M = 3.84–3.96 on a 5-point scale; Table 5). Usefulness (M = 3.96) and Ease of Understanding (M = 3.95) were rated highest, while Personal Relevance (M = 3.85) and Trust & Adoption (M = 3.84) were also favorable. Engagement indicators suggested sustained interaction (mean session duration = 12.6 minutes; mean follow-up turns = 4.2), with distributions shown in Figure 3.

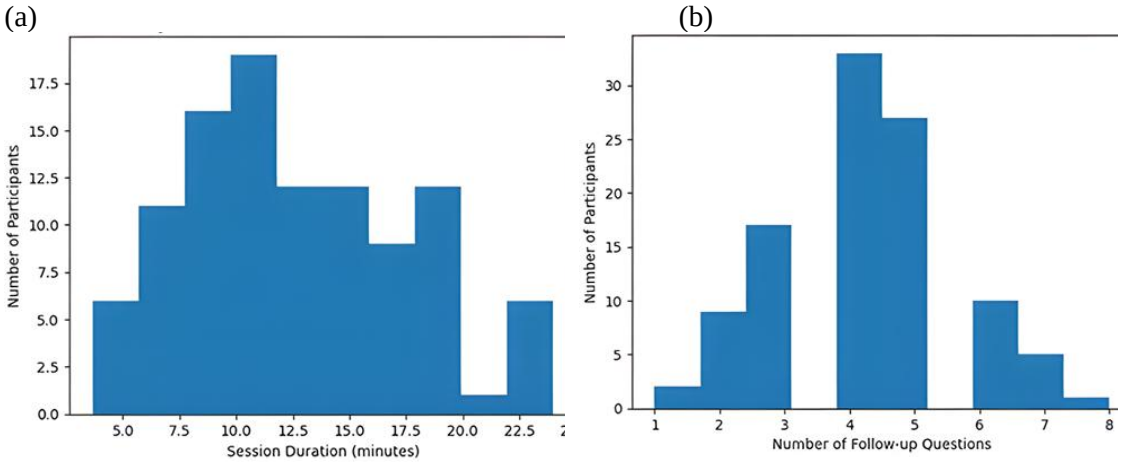


Figure 3. Engagement distribution metrics. (a) Distribution of session duration, showing anaverage interaction time of approximately 12.6 minutes, indicating sustained user engagement with the LLM-based consultation service. (b) Distribution of follow-up questions, with most participants initiating 3–5 additional turns, reflecting active and iterative interaction rather than passive information consumption.

Table 5 Descriptive statistics for User Experience (UX) constructs (N = 104).

UX Construct	Mean	SD	Median	Interpretation
Perceived Usefulness	3.96	0.82	4.0	High
Ease of Understanding	3.95	0.75	4.0	High
Personal Relevance	3.85	0.86	4.0	Positive
Trust & Adoption	3.84	0.77	4.0	Positive

Note: All items measured on a 5-point Likert scale (1=Strongly Disagree, 5=Strongly Agree).

Pre–Post Security Behavior Intentions (RQ3)

Paired-samples tests indicated a statistically significant pre–post increase in total SeBIS scores observed after the consultation session (Mpre = 3.67, SD = 0.59; Mpost = 3.96, SD = 0.61), $t(103) = 4.32$, $p < .001$, with a medium within-subject effect (dz = 0.42). Subscale analyses showed significant gains across all dimensions, with the largest improvement in Password Management (dz = 0.51) and smaller

gains for Device Security (dz = 0.23), consistent with a higher baseline and reduced headroom (Table 6). Because difference scores deviated from normality, Wilcoxon signed-rank tests were used as robustness checks and yielded convergent conclusions after Holm adjustment.

Table 6 Comparison of pre- and post-intervention SeBIS scores (N=104)

Construct	Pre-test Mean (SD)	Post-test Mean (SD)	t(103)	p	Cohen's $d_{(z)}$ *
Total SeBIS	3.67 (0.59)	3.96 (0.61)	4.32	< .001	0.42
Password Management	3.36 (1.06)	3.87 (0.80)	5.23	< .001	0.51
System Updates	3.68 (0.93)	3.94 (0.84)	2.93	0.004	0.29
Proactive Awareness	3.75 (0.83)	3.98 (0.74)	2.61	0.010	0.26
Device Security	3.88 (0.77)	4.06 (0.64)	2.39	0.019	0.23

Note. $d_{(z)}$ represents Cohen's d for paired samples (0.20 = small, 0.50 = medium, 0.80 = large).

Associations Between UX/Engagement and ΔSeBIS (Exploratory)

Spearman correlations (two-tailed) indicated that ΔSeBIS (SeBIS_post − SeBIS_pre) was positively associated with all UX constructs and overall UX (Figure 4). The strongest association was Personal Relevance (r_s=0.457), followed by Overall UX (r_s=0.398) and session duration (r_s=0.384); follow-up turns also correlated with ΔSeBIS (r_s=0.328). Profile variables showed no significant associations with ΔSeBIS across gender, age group, education, field of study, daily internet use, prior training source, self-rated knowledge, or incident experience (all p>.05); daily internet use showed a weak non-significant trend (Kruskal–Wallis p=.074). These analyses are exploratory and do not imply causality.

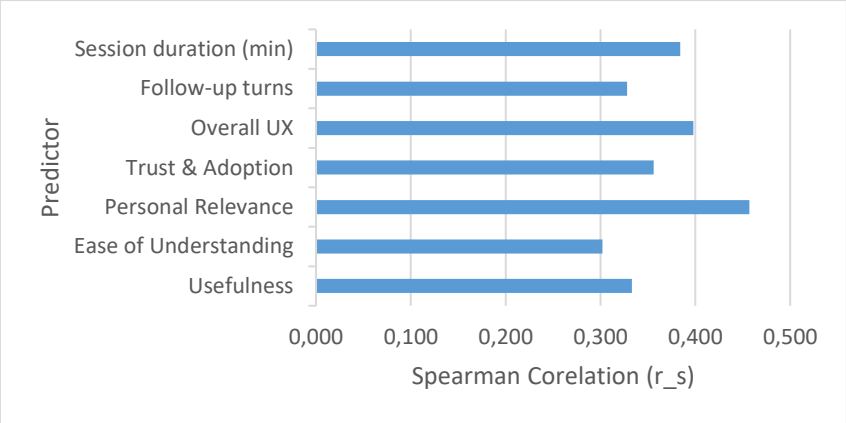


Figure 4 Spearman correlations between UX/engagement indices and ΔSeBIS(n = 104)

From national benchmarking to "just-in-time" personalized remediation

Findings. Participants showed a statistically significant improvement in overall security behavior intentions after the LLM-based consultation (SeBIS total: Mpre = 3.67 → Mpost = 3.96; dz ≈ 0.42), alongside consistently positive UX ratings (all constructs ~4/5). The improvement pattern was robust to non-parametric re-testing (Wilcoxon), suggesting the finding is not driven by normality assumptions.

Interpretation & comparison. The magnitude and direction of this change are consistent with what the ISA literature often reports for single-session interventions: immediate gains are frequently observed, but durable transfer typically benefits from reinforcement and repeated practice rather than one-off exposure (Prümmer et al., 2024; Schöni et al., 2025). Importantly, positioning the observed effect within the broader evidence base, a recent meta-analysis of end-user cybersecurity training reports a medium-to-large overall impact (d ≈ 0.75), with substantially larger effects on predictors of behavior (e.g., attitudes/knowledge; d ≈ 1.02) than on behavior change itself (d ≈ 0.36, often non-significant) (Prümmer et al., 2025). Because the present study measures behavior intentions (SeBIS) as a proximal outcome rather than directly observed behaviors, an effect in the moderate range (dz ≈ 0.42) is plausible and aligns with the broader pattern that antecedents/precursors of secure behavior are more readily shifted than

behavior itself. Within a Protection Motivation Theory (PMT) lens, this pattern is also expected: awareness gains do not reliably translate into action unless the intervention strengthens coping appraisal (response efficacy and self-efficacy) and lowers response costs (Khan et al., 2023). The LLM-based consultation can be interpreted as an operational mechanism for coping appraisal because it converts abstract assessment outputs into prioritized, concrete, low-friction actions that users can enact immediately (Parsons et al., 2017).

Implications. Theoretically, these results provide applied evidence that coping appraisal can be operationalized at scale through conversational personalization, suggesting a pathway by which national awareness instruments can move beyond KAB-style "measurement of awareness" toward PMT-style action enablement. Practically, this supports designing survey infrastructures as closed action loops—assessment → weakness diagnosis → prioritized plan → verification checklist—while treating SeBIS change as an immediate proximal outcome that can be complemented by retention follow-ups in later phases. For a national system, the key design claim is not that "LLMs educate," but that LLMs can reliably translate abstract scores into actionable micro-interventions when motivation is at its peak.

Why procedural behaviors improve more than situational vigilance

Findings. Gains were not uniform: Password Management improved most ($dz \approx 0.51$), while Device Security improved least ($dz \approx 0.23$) from a relatively high baseline (ceiling tendency). Proactive Awareness and System Updates showed modest but significant gains. Interpretation & comparison. This pattern aligns with a practical distinction between procedural behaviors (discrete steps that can be "implemented" immediately) and situational vigilance (judgment under uncertainty). Password behaviors are readily translated into concrete actions (e.g., enabling 2FA, adopting a password manager), whereas proactive vigilance against phishing/social engineering requires recognition skill, exposure to diverse examples, and corrective feedback—capabilities that are harder to cultivate in a brief, single-session interaction (Jampen et al., 2020).

Evidence syntheses in usable security emphasize that phishing resistance is strongly method-dependent and often benefits from practice-oriented designs (e.g., simulations, repeated exercises) and feedback loops rather than one-off informational messaging (Hillman et al., 2023; Marshall et al., 2024). Field evidence further suggests that awareness decays over time and that timed reminders/booster interventions can be necessary to sustain recognition and reduce susceptibility, reinforcing the need for reinforcement when the target is vigilance rather than procedural compliance (Reinheimer et al., 2020). In contrast, the smaller Device Security gain is consistent with a limited headroom effect: when baseline Likert ratings are already high, the measurable room for improvement is constrained even if the intervention is helpful. Finally, psychometric work on smartphone security underscores that security behavior intentions are multidimensional and context-sensitive, supporting the expectation that short interventions may shift some dimensions (procedural/tool-adoption) more readily than others (contextual detection and judgment) (Alanazi et al., 2022; Huang et al., 2023).

Implications. Theoretically, the subscale gradient suggests conversational micro-interventions are most efficient for targets that can be proceduralized into short checklists and tool adoption, while vigilance-related constructs likely require practice and reinforcement. Practically, to raise Proactive Awareness within the SKKS+consultation workflow, the consultation layer can be augmented with (i) micro-drills using personalized scam examples, (ii) simple heuristics (e.g., "pause–verify–report"), (iii) short self-tests with immediate feedback, and (iv) booster prompts (spaced reminders) that re-activate recognition skills over time—without requiring a full standalone training program (Hong & Furnell, 2021).

Personal relevance and interactive engagement as likely mechanisms of change

Findings. Personal Relevance correlated more strongly with improvement ($rs \approx 0.46$) than Ease of Understanding ($rs \approx 0.30$), and engagement indicators suggested active use (multi-turn follow-ups $rs \approx 0.33$; sustained session duration $rs \approx 0.38$). Profile variables showed no statistically significant association with ΔSeBIS (all $p > .05$), suggesting the short-term intention gains were not concentrated in a specific subgroup within the sampled Gen-Z population.

Interpretation & comparison. This pattern suggests that comprehension is not the primary bottleneck; rather, the key constraint is contextual fit—whether guidance maps onto users' routines, constraints, and threat exposures. This aligns with personalization arguments in security awareness that effective messaging should reflect user differences and situational realities, not merely simplify language

(Alotaibi et al., 2023; McCormac et al., 2017). In parallel, meta-analytic evidence on LLM persuasion emphasizes substantial heterogeneity across contexts and indicates that implementation choices—such as personalization and interactive (multi-turn) delivery—are plausible moderators of downstream impact, even when individual moderators may not always reach significance in small evidence bases (Hölbling et al., 2025). Taken together, the stronger relevance–gain association and the observed engagement behavior are consistent with an interpretation that users benefit when they can iteratively clarify, personalize, and "translate" advice into their own context during dialogue (Hillman et al., 2023).

Implications. For practice, the results motivate designing cybersecurity education around profile-driven personalization plus interaction scaffolds, rather than further optimizing generic readability. LLMs are increasingly positioned as enablers for this at scale because they can transform individual risk profiles into tailored, actionable guidance and sustain interactive coaching that may improve engagement and contextual fit compared with static advice (Sun et al., 2025; Xu et al., 2025). Concretely, the consultation layer can be strengthened by (i) prompted follow-ups (suggested questions based on the user's weaknesses), (ii) teach-back checks ("What will you change first?"), (iii) micro-quizzes/self-tests tied to the user's risk profile, and (iv) action summarization (a short and prioritized checklist at exit). Finally, these correlational findings should be framed as exploratory and mechanism-suggestive (not causal), but they provide actionable design signals about where personalization is likely to matter most.

Trust is beneficial—but must be paired with calibrated reliance and safety-by-design

Findings. Users reported high Trust & Adoption intentions, while the expert panel rated Safety & Ethics very highly and Accuracy slightly lower than the top dimensions. This combination is desirable for uptake, but it raises a familiar risk in human–AI systems: trust can outpace correctness, especially when outputs are fluent and confident.

Interpretation & comparison. This pattern is consistent with the broader literature that treats trust as a multi-faceted and inconsistently operationalized construct—one that is often studied cross-sectionally and can be difficult to calibrate over time (Ng & Zhang, 2025). In practice, higher trust can translate into higher reliance behavior: experimental evidence shows that participants may follow AI advice even when conflicting contextual information is available, with measurable costs from overreliance (Klingbeil et al., 2024). For generative systems, the primary risk is not only overtly unsafe or malicious content, but also plausible-yet-suboptimal guidance that is accepted because it "sounds right"—a failure mode explicitly captured in LLM risk taxonomies (misleading information, downstream misuse, and overreliance) even when content is not overtly unsafe. Research syntheses on appropriate reliance therefore argue that product success depends on helping users accept correct outputs while rejecting incorrect ones, and that calibration requires both interface design and policy/guardrail mechanisms rather than relying on user skepticism alone (Passi et al., 2024). Importantly, transparency mechanisms are mixed: explanations can reduce overreliance when they lower verification costs and support selective scrutiny (Vasconcelos et al., 2023), yet misleading or low-quality explanations can create "halo" effects that increase incorrect acceptance—or even teach users flawed reasoning patterns—despite correct recommendations (Cabitza et al., 2024).

Why dual-layer evaluation strengthens claims beyond "the model looks good"

Findings. The study combined (i) model-centric expert ratings with strong agreement and (ii) user-centric evidence of perceived usefulness and measurable pre–post improvement. Interpretation & comparison. Human evaluation frameworks argue that single-score evaluations are insufficient because LLM quality is multidimensional (relevance, correctness, clarity, safety) and context-dependent. ConSiDERS explicitly frames human evaluation as multidisciplinary and warns against narrow metrics that miss user experience and real-world constraints (Elangovan et al., 2024). HumanELY similarly emphasizes structured, rubric-based human evaluation to make judgments more comparable and auditable (Awasthi et al., 2023). The present results fit these recommendations by demonstrating that strong safety/quality signals in expert review can coincide with high user trust and measurable behavioral-intention change—supporting a more defensible "system works end-to-end" claim than model-only evaluation (Es et al., 2024).

Theoretically, this study supports a socio-technical evaluation stance: LLM systems should be judged by end-to-end human outcomes, not only textual quality. Practically, public-facing cybersecurity deployments benefit from an evaluation stack that (i) gates deployment with expert safety/quality review, and (ii) monitors user-side outcomes (relevance, trust calibration, and behavior intent change). This

structure is defensible for safety-critical domains and provides a repeatable template for scaling AI-assisted awareness programs.

Limitations and Future Research

It is important to clarify that the present study evaluates changes in security behavior intentions rather than objectively verified security behaviors. Within behavioral security research, intention is widely treated as a proximal predictor of behavior, yet it does not guarantee sustained real-world adoption. The observed improvements therefore indicate enhanced motivational readiness and coping appraisal, but future studies incorporating objective behavioral indicators (e.g., 2FA activation logs, phishing simulation outcomes) are necessary to confirm durable behavior change.

This study employed a single-session within-subject pre–post design without a randomized control group. Consequently, observed improvements in SeBIS intentions cannot be attributed exclusively to the LLM-based consultation and may partly reflect testing effects or short-term motivational boosts, a limitation commonly noted in ISA interventions (Prümmer et al., 2025). Second, outcomes relied on self-reported behavior intentions rather than objectively verified behaviors. While SeBIS is a validated proximal indicator, intention does not always translate into sustained real-world practice. Third, the evaluation captured only immediate post-intervention effects. Prior work on phishing and vigilance training indicates that durable behavior change typically requires repeated practice and reinforcement, suggesting that short-term gains may not persist without follow-up. Finally, the sample reflects a specific demographic and national context (Gen Z users in Indonesia), which may limit generalizability. The study also relied on a non-probability convenience sampling strategy, resulting in a youth-heavy participant pool. While appropriate for an exploratory pilot within the national SKKS context, this sampling approach may limit external validity across older populations or users with substantially different digital literacy profiles. The combination of high user trust with nontrivial accuracy variance also highlights a known sociotechnical risk: systems can appear safe yet still elicit overreliance on confidently delivered but suboptimal advice (Weidinger et al., 2022), underscoring the need for calibrated trust.

Because both user experience constructs and SeBIS scores were collected via self-report within the same session, shared method variance and demand characteristics may partially inflate observed correlations. Although the within-subject design reduces between-participant variability, future research should incorporate objective behavioral measures, delayed follow-up assessments, or experimental control conditions to better isolate causal mechanisms and reduce common-method bias. While the 20-profile test set was designed to maximize structural variation across SKKS maturity bands and vulnerability types, it does not exhaustively represent all possible edge cases or adversarial prompts. Future work should incorporate adversarial robustness testing and injection-resistant evaluation scenarios

Future work should prioritize causal designs, such as randomized or A/B-controlled studies comparing personalized interactive consultation, static advice, and assessment-only feedback, to isolate the active components of effectiveness. Longitudinal studies are needed to assess retention and behavior transfer, ideally incorporating objective behavioral indicators (e.g., 2FA adoption, phishing simulation performance) rather than relying solely on self-report. Given the weaker gains in vigilance-related constructs, future systems should embed micro-practice mechanisms—personalized examples, short quizzes, and immediate feedback—to better support situational awareness. Finally, trust calibration and appropriate reliance should be treated as primary outcomes alongside UX. Design experiments should evaluate grounding cues, uncertainty signaling, and verification scaffolds, while accounting for mixed evidence regarding explanations, which may either reduce or exacerbate overreliance depending on context (Spitzer et al., 2025). Continuous expert evaluation and adversarial testing, as recommended by human-centered evaluation frameworks, will be essential for scaling deployment in safety-critical public systems.

CONCLUSION

This study integrated an LLM-based consultation service into the SKKS workflow to translate national-scale awareness benchmarking into immediate, personalized remediation. Using a two-phase evaluation, expert review indicated that the system produced high-quality recommendations with strong safety performance and reliable agreement across raters, supporting initial feasibility for controlled public-facing evaluation. In the user study, participants reported consistently positive experience and

sustained engagement, and they showed a statistically significant pre–post improvement in security behavior intentions as measured by SeBIS, with the largest gains in procedural domains such as password management. Exploratory analyses further suggested that perceived personal relevance and interactive engagement were positively associated with intention gains, highlighting personalization as a plausible mechanism. Overall, the findings suggest that LLM-mediated, profile-driven consultation shows promise as an adaptive remediation layer within national cybersecurity awareness infrastructures. While short-term intention gains should not be conflated with verified long-term behavior change, the dual-layer evaluation provides convergent evidence of safety, perceived usefulness, and measurable motivational improvement. When combined with safety-by-design principles and ongoing human-centered evaluation, such systems offer a scalable pathway for transforming benchmarking instruments into responsive behavioral intervention ecosystems. Practically, the findings support embedding profile-driven, safety-by-design consultation layers into public cybersecurity programs to deliver actionable guidance at scale. While effects reflect short-term intention change rather than verified long-term behavior, the results provide a defensible foundation for developing AI-assisted awareness systems that are both responsive and responsibly governed.

ACKNOWLEDGMENTS

This work was supported by BSSN under Grant number (759/PS/TA.01/02/02/2026).

AUTHOR CONTRIBUTIONS

Conceptualization, R.B.H.; Methodology, R.B.H. and R.D.; Software, R.D.; Investigation, R.D.; Data Curation, R.D.; Writing—Original Draft Preparation, R.B.H.; Writing—Review & Editing, N.Q.; Supervision, R.B.H.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

USE OF ARTIFICIAL INTELLIGENCE (AI)-ASSISTED TECHNOLOGY

The authors used an AI-based language model solely to enhance the grammar, language clarity, and overall readability of this manuscript. All content, analyses, interpretations, and conclusions were developed by the authors. The manuscript has been carefully reviewed, revised, and approved by the authors to ensure accuracy, integrity, and full responsibility for its content.

REFERENCES

Abduel, M. (2024). Users' awareness of cyber security practices for preventing data attacks in public organisations. *The Journal of Informatics*, *4*(1). https://doi.org/10.59645/tji.v4i1.355

Ahmad, M. R., Osman, M. H., Abdullah, A., & Sharif, K. Y. (2024). Evolution of information security awareness towards maturity: A systematic review. *International Journal on Advanced Science, Engineering and Information Technology*, *14*(5), 1738–1747. https://doi.org/10.18517/ijaseit.14.5.20234

Alanazi, M., Freeman, M., & Tootell, H. (2022). Exploring the factors that influence the cybersecurity behaviors of young adults. *Computers in Human Behavior*, *136*, 107376. https://doi.org/10.1016/j.chb.2022.107376

Alotaibi, S., Furnell, S., & He, Y. (2023). Towards a framework for the personalization of cybersecurity awareness. in s. furnell & n. clarke (Eds.), *Human Aspects of Information Security and Assurance* (Vol. 674, pp. 143–153). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-38530-8_12

Alrababah, H., Iqbal, H., & Khan, M. A. (2024). The effect of user behavior in online banking on cybersecurity knowledge. *International Journal of Intelligent Systems*, *2024*(1), 9949510. https://doi.org/10.1155/int/9949510

Awasthi, R., Mishra, S., Mahapatra, D., Khanna, A., Maheshwari, K., Cywinski, J., Papay, F., & Mathur, P. (2023). Human ELY: Human evaluation of LLM yield, using a novel web-based evaluation tool. *Health Informatics.* https://doi.org/10.1101/2023.12.22.23300458

Ayyagari, R. (2020). Risk and demographics' influence on security behavior intentions. *Journal of the Southern Association for Information Systems*, *7*(1). https://doi.org/10.17705/3JSIS.00013

BSSN. (2025). *Cyber Security Awareness Survey (SKKS) Report 2024*. Badan siber dan sandi Negara. https://www.bssn.go.id/wp-content/uploads/2025/02/Laporan-SKKS-2024.pdf

Cabitza, F., Fregosi, C., Campagner, A., & Natali, C. (2024). Explanations considered harmful: The impact of misleading explanations on accuracy in hybrid human-ai decision making, *Explainable Artificial Intelligence, 21(4)* pp. 255–269). https://doi.org/10.1007/978-3-031-63803-9_14

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

Egelman, S., Harbach, M., & Peer, E. (2016). Behavior ever follows intention? a validation of the security behavior intentions scale (SeBIS). *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16*, 5257–5261. https://doi.org/10.1145/2858036.2858265

Elangovan, A., Liu, L., Xu, L., Bodapati, S. B., & Roth, D. (2024). Considers-the-human evaluation framework: rethinking human evaluation for generative large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1137–1160. https://doi.org/10.18653/v1/2024.acl-long.63

Es, S., James, J., Espinosa Anke, L., & Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 150–158). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.eacl-demo.16

Gessinger, I., Seaborn, K., Steeds, M., & Cowan, B. R. (2025). ChatGPT and me: First-time and experienced users' perceptions of ChatGPT's communicative ability as a dialogue partner. *International Journal of Human-Computer Studies*, *194*, 103400. https://doi.org/10.1016/j.ijhcs.2024.103400

Hillman, D., Harel, Y., & Toch, E. (2023). Evaluating organizational phishing awareness training on an enterprise scale. *Computers & Security*, *132*, 103364. https://doi.org/10.1016/j.cose.2023.103364

Hölbling, L., Maier, S., & Feuerriegel, S. (2025). A meta-analysis of the persuasive power of large language models. *Scientific Reports*, *15*(1), 43818. https://doi.org/10.1038/s41598-025-30783-y

Hong, Y., & Furnell, S. (2021). Understanding cybersecurity behavioral habits: Insights from situational support. *Journal of Information Security and Applications*, *57*, 102710. https://doi.org/10.1016/j.jisa.2020.102710

Huang, H.-Y., Demetriou, S., Hassan, M., Tuncay, G. S., Gunter, C. A., & Bashir, M. (2023). Evaluating User Behavior in Smartphone Security: A Psychometric Perspective. *Proceedings of the Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*, 509–524. https://www.usenix.org/conference/soups2023/presentation/huang

Iskandar, A. S., Hilman, M., & Yazid, S. (2026). Information security awareness assessment for civil servant recruitment committee in Indonesia using HAIS-Q. *Information & Computer Security*, *34*(1), 86–103. https://doi.org/10.1108/ICS-01-2025-0019

Jampen, D., Gür, G., Sutter, T., & Tellenbach, B. (2020). Don't click: Towards an effective anti-phishing training. A comparative literature review. *Human-Centric Computing and Information Sciences*, *10*(1), 33. https://doi.org/10.1186/s13673-020-00237-7

Khan, N. F., Ikram, N., Murtaza, H., & Javed, M. (2023). Evaluating protection motivation based cybersecurity awareness training on Kirkpatrick's Model. *Computers & Security*, *125*, 103049. https://doi.org/10.1016/j.cose.2022.103049

Kim, J., Kim, J. H., Kim, C., & Park, J. (2023). Decisions with ChatGPT: Reexamining choice overload in ChatGPT recommendations. *Journal of Retailing and Consumer Services*, *75*, 103494. https://doi.org/10.1016/j.jretconser.2023.103494

Klingbeil, A., Grützner, C., & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Kurniawan, Y., Santoso, S. I., Wibowo, R. R., Anwar, N., Bhutkar, G., & Halim, E. (2023). Analysis of Higher Education Students' awareness in Indonesia on personal data security in social media. *Sustainability*, *15*(4), 3814. https://doi.org/10.3390/su15043814

Li, L., Zhang, Y., & Chen, L. (2023). Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, *41*(4), 1–26. https://doi.org/10.1145/3580488

Manzoor, A., Ziegler, S. C., Garcia, K. Maria. P., & Jannach, D. (2024). ChatGPT as a conversational recommender system: A user-centric analysis. *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 267–272. https://doi.org/10.1145/3627043.3659574

Marshall, N., Sturman, D., & Auton, J. C. (2024). Exploring the evidence for email phishing training: A scoping review. *Computers & Security*, *139*, 103695. https://doi.org/10.1016/j.cose.2023.103695

McCormac, A., Zwaans, T., Parsons, K., Calic, D., Butavicius, M., & Pattinson, M. (2017). Individual differences and information security awareness. *Computers in Human Behavior*, *69*, 151–156. https://doi.org/10.1016/j.chb.2016.11.065

Mehrotra, S., Degachi, C., Vereschak, O., Jonker, C. M., & Tielman, M. L. (2024). A systematic review on fostering appropriate trust in human-ai interaction: trends, opportunities and challenges. *ACM J. Responsib. Comput.*, *1*(4). https://doi.org/10.1145/3696449

Ng, S. W. T., & Zhang, R. (2025). Trust in AI chatbots: A systematic review. *Telematics and Informatics*, *97*, 102240. https://doi.org/10.1016/j.tele.2025.102240

Noh, H.-H., Rim, H. B., & Lee, B.-K. (2025). Exploring user attitudes and trust toward chatgpt as a social actor: A utaut-based analysis. *Sage Open*, *15*(2), 21582440251345896. https://doi.org/10.1177/21582440251345896

Pandey, S., & Sharma, S. (2023). A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning. *Healthcare Analytics*, *3*, 100198. https://doi.org/10.1016/j.health.2023.100198

Parsons, K., Calic, D., Pattinson, M., Butavicius, M., McCormac, A., & Zwaans, T. (2017). The human aspects of information security questionnaire (HAIS-Q). *Comput. Secur.*, *66*(C), 40–51. https://doi.org/10.1016/j.cose.2017.01.004

Passi, S., Dhanorkar, S., & Vorvoreanu, M. (2024). *Appropriate Reliance on Generative AI: Research Synthesis*. Microsoft Research.

Perkasa, D. A., & Setiawan, B. (2024). Measuring information security awareness level of high school students. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, *4*(4), 1301–1308. https://doi.org/10.57152/malcom.v4i4.1461

Prümmer, J., Van Steen, T., & Van Den Berg, B. (2024). A systematic review of current cybersecurity training methods. *Computers & Security*, *136*, 103585. https://doi.org/10.1016/j.cose.2023.103585

Prümmer, J., Van Steen, T., & Van Den Berg, B. (2025). Assessing the effect of cybersecurity training on End-users: A Meta-analysis. *Computers & Security*, *150*, 104206. https://doi.org/10.1016/j.cose.2024.104206

Reinheimer, B., Aldag, L., Mayer, P., Mossano, M., Duezguen, R., & Volkamer, M. (2020). An Investigation of phishing awareness and education over time: when and how to best remind users. *Proceedings of the Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. https://www.usenix.org/conference/soups2020/presentation/reinheimer

Rizal, M. A., & Setiawan, B. (2024). Information security awareness literature review: focus area for measurement instruments. *Procedia Computer Science*, *234*, 1420–1427. https://doi.org/10.1016/j.procs.2024.03.141

Robbins, M. S., & Robbins, C. (2025). Impact of information security awareness training on knowledge, attitude, and behavior: A K-12 Case Study. *Journal of Cybersecurity Education, Research and Practice*, *2025*(1). https://doi.org/10.62915/2472-2707.1252

Sawaya, Y., Lu, S., Isohara, T., & Sharif, M. (2024). A high coverage cybersecurity scale predictive of user behavior. *33rd USENIX Security Symposium (USENIX Security 24)*, 5503–5520. https://www.usenix.org/conference/usenixsecurity24/presentation/sawaya

Schöni, L., Roch, N., Sievers, H., Strohmeier, M., Mayer, P., & Zimmermann, V. (2025). It's a Match—Enhancing the fit between users and phishing training through personalisation. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–25. https://doi.org/10.1145/3706598.3713845

Shari, A. M. J., Ahmad, M., Razali, R. R. R., & Sujak, A. F. A. (2023). Knowledge, attitude, and practices towards internet safety and security among generation Z in Malaysia: A conceptual paper, *Proceedings of the International Conference on Communication, Language, Education and Social Sciences (CLESS 2022)* (pp. 4–10). Atlantis Press SARL. https://doi.org/10.2991/978-2-494069-61-9_2

Shelby, R., Rismani, S., Henne, K., Moon, Aj., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N., Gallegos, J., Smart, A., Garcia, E., & Virk, G. (2023). Sociotechnical harms of algorithmic systems: scoping a taxonomy for harm reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741. https://doi.org/10.1145/3600211.3604673

Spitzer, P., Holstein, J., Morrison, K., Holstein, K., Satzger, G., & Kühl, N. (2025). Don't Be Fooled: The misinformation effect of explanations in human–AI collaboration. *International Journal of Human–Computer Interaction*, 1–29. https://doi.org/10.1080/10447318.2025.2574511

Sun, N., Miao, Y., Mo, X., & Zhang, J. (2025). Large language models for cybersecurity education: a survey of current practices and future directions. *Data Science: Foundations and Applications* (Vol. 15875, pp. 3–20). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-8295-9_1

Taherdoost, H. (2024). A critical review on cybersecurity awareness frameworks and training models. *Procedia Computer Science*, *235*, 1649–1663. https://doi.org/10.1016/j.procs.2024.04.156

Tam, T. Y. C., Sivarajkumar, S., Kapoor, S., Stolyar, A. V., Polanska, K., McCarthy, K. R., Osterhoudt, H., Wu, X., Visweswaran, S., Fu, S., Mathur, P., Cacciamani, G. E., Sun, C., Peng, Y., & Wang, Y. (2024). A framework for human evaluation of large language models in healthcare derived from literature review. *Npj Digital Medicine*, *7*(1), 258. https://doi.org/10.1038/s41746-024-01258-7

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38. https://doi.org/10.1145/3579605

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of risks posed by language models. *2022 ACM Conference on Fairness Accountability and Transparency*, 214–229. https://doi.org/10.1145/3531146.3533088

Xu, H., Wang, S., Li, N., Wang, K., Zhao, Y., Chen, K., Yu, T., Liu, Y., & Wang, H. (2025). Large language models for cyber security: A systematic literature review. *ACM Transactions on Software Engineering and Methodology*, 3769676. https://doi.org/10.1145/3769676