National College of Ireland

# Smishlock Holmes
# Explainable Smishing Detector

MSc Research Project
MSc Cybersecurity

Andre Luis Oppenheimer Marques
Student ID: 23355280

School of Computing
National College of Ireland

Supervisor: Joel Aleburu

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student Name:** ANDRE LUIS OPPENHEIMER MARQUES

**Student ID:** 23355280

**Programme:** MSCCYB **Year:** 2025

**Module:** PRACTICUM

**Supervisor:**

**Submission Due Date:**

**Project Title:** Smishlock Holmes

**Word Count:** 5985 **Page Count** 17

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Andre Luis Oppenheimer Marques

**Date:** 11/08/2025

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Smishlock Holmes
# Explainable Smishing Detector

Andre Luis Oppenheimer Marques
23355280

**Abstract**

Smishing (SMS phishing) is a rapidly evolving social engineering threat that exploits both technical evasion tactics and cognitive vulnerabilities in mobile-first environments. Traditional defences often fail against adversarially crafted messages and rarely provide explanations that enhance user awareness and trust. This study investigates how Large Language Models (LLMs) can be applied to detect smishing messages in a way that combines classification robustness with interpretable, context-aware explanations for cognitively constrained mobile users.

We present **Smishlock Holmes**, a modular AI framework that detects technical, psychological, and exploitation cues, normalises obfuscated text, and produces causal narratives linking cues to risk. Built on a JSON-Embedded Few-Shot Chain-of-Thought architecture, it enforces schema-validated, model-agnostic outputs.

Using a balanced subset of the Super SMS Dataset, classification accuracy and explanation quality were assessed, the latter via the G-Eval framework. The system achieved **87.5% accuracy**, with high correctness (4.68/5) and clarity (4.50/5), but lower completeness (3.94/5).

These results show that LLM-based detection can unify adversarial resilience and interpretability in a single pipeline. Future work will expand cue coverage, adopt multi-agent verification, and explore privacy-preserving on-device inference for scalable, user-adaptive smishing defences.

# 1 Introduction

Smishing (SMS phishing) represents a well-established yet dynamically evolving cybersecurity threat, wherein malicious actors leverage mobile messaging channels to deceive users into divulging sensitive information or installing malware. As mobile devices have become integral to both personal and professional activities, attackers increasingly exploit these platforms to circumvent traditional email-based defenses. Recent industry analyses confirm that smishing has become one of the most pervasive forms of social engineering, sustained by its high success rates and users' inherent trust in SMS communications. According to Proofpoint's State of the Phish 2024 report, threat actors are diversifying their delivery vectors by exploiting urgency and brand impersonation across SMS and messaging platforms, taking advantage of users' reduced scrutiny on mobile devices. The widespread adoption of SMS-based authentication mechanisms, the prevalence of mobile-first digital services, and the normalization of logistics-related messaging have collectively created an environment in which smishing can thrive. High-impact incidents involving the impersonation of financial institutions, courier services, and government agencies have resulted in substantial financial and reputational harm, underscoring the need for robust, user-centred detection mechanisms.

Current smishing campaigns leverages of a potent blend of cognitive and technical vulnerabilities, attackers increasingly employ evasive obfuscation techniques like spacing attacks (inserting zero-width spaces or breaks in malicious words), homoglyph substitutions (replacing characters with visually similar ones), and Punycode URL encoding to disguise malicious links. Such techniques enable smishing messages to slip past both legacy keyword filters and many machine learning spam detectors that were not trained on these adversarial examples.Additionally, smishing success is not solely due to technical evasion, but also to deep-rooted psychological and behavioral factors. Users often rely on surface-level cues, such as message tone, urgency, or contextual-familiarity, while ignoring critical metadata. This heuristic-driven decision-making leaves even well-intentioned users vulnerable, particularly on mobile devices where cognitive bandwidth is limited. Together, these insights emphasize that technical safeguards alone are insufficient. Effective smishing detection must integrate explainable, user-centred strategies that interrupt automatic responses and promote informed reflection.

This leads to the core research question guiding the present study:
*How can Large Language Models (LLMs) be applied to detect smishing messages in a way that not only improves classification robustness but also generates interpretable explanations that enhance user awareness and trust, particularly in cognitively constrained mobile contexts?*

To address this question, the study introduces Smishlock Holmes, a generative AI agent that combines adversarial robust classification with layered, context-aware explanations tailored to the mobile user experience. Unlike traditional detectors, Smishlock Holmes highlights semantic risk cues, such as urgency framing, contextual familiarity, and obfuscated URLs, in natural language outputs understandable even to non-expert users. Classification accuracy will be validated using benchmark testing on gold-labeled balanced subset of Super SMS dataset, while explanation quality will be assessed using the peer-reviewed G-EVAL framework. This approach unites technical robustness with psychological usability, treating smishing defense not as a binary judgment engine, but as a collaborative cognitive interface that educates and supports users at the point of decision. However, several important limitations define the scope of this investigation.

First, smishing classification often involves gray-area messages, texts that cannot be definitively labeled as benign or malicious without access to sender identity or additional context. Smishlock Holmes explicitly acknowledges this ambiguity by shifting from rigid verdicts to risk-based interpretability. Second, the reliance on cloud-based LLMs raises privacy and regulatory considerations, particularly under GDPR. To mitigate this, the study confines its scope to public, non-personal data (the Super SMS Dataset), while future work will explore on-device inference or privacy-preserving techniques.

# 2 Related Work

## 2.1 State of Art Solution

Salman et al. (2024) and Seo et al. (2024) both address the vulnerability of smishing detectors to textual adversarial attacks, but their findings diverge sharply on what model architecture offers the most robust defense. While Salman supports transformer-based models like RoBERTa for their superior contextual resilience, Seo advocates for a lightweight Char-CNN

as a better defense against character-level evasion. This contrast reflects differences in dataset language, deployment goals, and model constraints.

Salman et al., using a newly released 67,000+ English SMS dataset, report that RoBERTa achieved a top F1-score of 98%, outperforming traditional models under both baseline and adversarial conditions. Their benchmark prioritizes large-scale comparability and adversarial stress testing. In contrast, Seo et al. train a 127kB Char-CNN on a proprietary Korean-language dataset augmented by their EVA variant generator. Despite its simplicity, the Char-CNN surpassed transformer models like KoELECTRA in both accuracy and evasion resistance under mobile constraints.

These differences may be partially explained by linguistic structure: Korean's agglutinative morphology and character-based script may benefit from character-level modeling, while English's syntax and semantic dependency on word context make transformer-based embeddings more effective. However, further cross-lingual adversarial analysis would be required to confirm this hypothesis. While both studies offer credible empirical support for their claims, Seo's lack of public data limits reproducibility. Salman's dataset and transparent benchmarks make their results more broadly applicable.

This work adopts the Salman et al. dataset as the core evaluation platform due to its language, size and public availability. Meanwhile, Seo et al. reinforce the importance of privacy-preserving and low-latency designs, proposing on-device inference as a practical deployment solution.

## 2.2 Why users feel for smishing

Timko et al. (2025) show that user interaction patterns with specific message elements, such as clicking on the sender or URL, significantly influence detection accuracy. In their study, clicking on the sender increased the odds of correctly identifying fake messages by 95.6% (Section 5). Behavioural metrics, including RSeBIS and IUIPC scores, were also associated with lower interaction rates with fraudulent content (Section 5.4). These results indicate that attention allocation is a key determinant of susceptibility, and that well-crafted cues can steer user focus away from diagnostic elements.

Complementing this, Tabassum et al. (2024) highlight that smishing susceptibility is not solely a function of message features, but also of the individual reasoning processes through which users assess legitimacy. Through qualitative interviews with 30 participants, they found that identical cues, such as urgency or personalization, were interpreted differently depending on prior experience and context. This subjectivity undermines static cue-based detection approaches and suggests that effective defences must adapt to varied cognitive frames.

Rahman et al. (2023) extend these insights with large-scale behavioural evidence from a U.S. telecom network, analysing over 368,000 smishing messages. They report that up to 10 % of recipients clicked malicious links, with engagement peaking during work hours and for financial-themed content. Notably, 2.9 % of users clicked within five minutes of message delivery, illustrating a behavioural impulsivity. This rapid response is frequently triggered by psychological levers—urgency, contextual familiarity, and perceived financial or social consequences—that are core features of smishing campaigns.

These behavioural and psychological vulnerabilities are amplified by the cognitive limitations of mobile device use. Figl and Remus (2023) found that smartphone users performed worse on Cognitive Reflection Tests than PC users, indicating a bias toward

intuitive, System 1 thinking over deliberative, System 2 reasoning. Liao et al. (2025) similarly showed, via eye-tracking, that mobile users had lower recall and attentional engagement when processing dense content, particularly under time pressure. Such constraints make it harder for users to engage with diagnostic cues even when they are noticed.

Taken together, these studies reveal a layered vulnerability in mobile smishing: attentional focus can be diverted, cue interpretation is subjective, impulsive responses are common, and mobile interfaces constrain reflective reasoning. Smishing exploits the convergence of these factors, succeeding not simply through deception, but by operating in an environment optimised for quick, surface-level decision-making.

## 2.3 The unstable relationship between trust and explanation

While explainability has become a central design goal in AI-human interaction, there is no stable consensus on how, or even whether, explanations foster user trust. Trust is variously framed as an epistemic state, a behavioural outcome, or a relational construct, leading to divergent assumptions and conflicting empirical results.

Baron (2025) challenges one of the field's foundational assumptions: that explainability is necessary for trust. He reframes trust as an agential rather than epistemic attitude, grounded in optimism and normative expectations rather than in transparent reasoning, and critiques canonical works such as Wachter et al. (2017) for conflating interpretability with trustworthiness. Although lacking empirical grounding, Baron's analysis is conceptually rigorous and directly relevant to Smishlock Holmes, which models trust not through exposure of internal model logic, but through contextualized, user-facing explanations and behavioural cues. In this framing, trust may arise from perceived competence and interaction quality, independent of full transparency.

Empirical evidence further reflects the instability of this relationship. Labarta et al. (2024) found that participants consistently preferred semantically rich explanations, such as counterfactuals and concept-based reasoning, over low-level saliency maps. Yet, these preferences did not consistently translate into improved decision-making or sustained behavioural trust.

Similarly, Papenmeier et al. (2022) demonstrate that neither higher model accuracy nor the mere presence of explanations guarantees greater trust. In a large-scale between-subjects experiment with 959 participants, the highest trust ratings emerged in a high-accuracy/no-explanation condition, while explanations sometimes reduced trust when misaligned with task complexity. This finding cautions against assuming that any explanation will be beneficial; in security contexts, mismatched detail or style could erode confidence even in accurate classifications.

Meta-analytic work by Atf and Lewis (2025) reinforces this caution. Across studies, explainability was positively but only weakly correlated with trust ($r = 0.194$), suggesting that transparency alone is insufficient for robust trust. They argue that explanation mechanisms must be embedded within a broader socio-technical framework that includes ethical safeguards, inclusivity, and user-identity considerations.

Chamola et al. (2023) further stress that explainability should be a means to enabling user oversight, contestability, and agency, not an end in itself. They catalogue a range of XAI techniques but note the gap between conceptual taxonomies and their practical deployment, particularly in adversarial or noisy environments.

Taken together, these studies reveal that the trust–explanation relationship is contingent, context-dependent, and easily undermined when explanation form, granularity, or intent

misalign with user expectations and cognitive constraints. For mobile smishing detection, this means that explanation strategies must be user-adaptive and context-aware, balancing conceptual clarity with behavioural resonance to support trust without overloading or alienating the user.

## 2.4 LLMs in Security Detection

The application of large language models (LLMs) to cybersecurity has accelerated rapidly, yet the literature reveals unresolved questions regarding optimal architecture, adversarial robustness, and the ethical deployment of such systems in security-critical contexts.

Chen et al. (2024) offer a layered taxonomy for LLM-based cyber threat detection, distinguishing backbone architectures, enhancement strategies, and target tasks. Within this framework, GPT-style decoder-only models are primarily positioned for generative functions, while classification is most often delegated to encoder-only architectures such as BERT. Smishlock Holmes departs from this separation, applying a GPT model to both classification and explanation. This unifies detection, adversarial text normalization, and user-facing interpretation within a single decoder-only backbone. While Chen et al. recognise few-shot and multi-task capabilities of GPTs, they do not fully explore this integrated pipeline design, which simplifies the architecture while maintaining interpretability through prompt-based control.

The performance advantages of LLMs in phishing detection are evident in work by Heiding et al. (2024), where GPT-3.5 achieved 84 % accuracy compared to 70 % for human participants. However, the authors note that internal variance persists, and the evaluated models lack mechanisms for transparent reasoning, raising concerns about trust calibration. This underscores a trade-off between statistical performance and explainability in high-stakes domains.

Evidence from Rashid et al. (2025) further validates the dual-task potential of LLMs. In their one-shot prompting experiments on malicious URL detection, GPT-4 achieved an F1-score of 96.2 % while producing coherent, informative explanations assessed via an independent rubric. Although the study acknowledges prompt sensitivity and limited language coverage, the results support Smishlock Holmes' design choice to pair classification and explanation within a single, adaptable prompting strategy.

Robustness to adversarial evasion remains a critical challenge. Salman et al. (2025) show that conventional ML models degrade sharply (8–30 % accuracy) against obfuscated SMS, whereas LLM-based approaches such as SpaLLM-Guard maintain performance up to 95 % under spacing, homoglyph, and leetspeak attacks. This resilience reflects the generalisation abilities of pre-trained language models. Nonetheless, their evaluation is limited to English-language datasets, with no assessment of privacy risks or model misuse. For Smishlock Holmes, these findings point to the value of prompt engineering and model pairing for adversarial robustness.

Across these studies, two key ethical considerations emerge. First, the opacity of LLM decision-making raises risks of over-reliance or misplaced trust if explanations are absent or poorly aligned with user mental models. Second, the integration of LLMs into messaging platforms introduces privacy concerns, particularly when sensitive communications must be processed for threat detection.

In summary, the current literature confirms that LLMs can surpass human baselines in phishing and smishing detection, generalise well to adversarial text perturbations, and—when prompted effectively—combine classification with interpretable explanations. However, these capabilities must be balanced against prompt sensitivity, language and domain generalisation limits, and the ethical imperatives of privacy, user trust, and responsible deployment.

# 3 Research Methodology

## 3.1 Threat Modeling

In adversarial settings, attackers attempt to bypass keyword-based filters and deceive users through subtle manipulations of textual content. Smishlock Holmes addresses this challenge through a two-step process: technical cue detection followed by normalization. First, messages are parsed for obfuscation tactics that typically evade automated detection while remaining readable to humans. These tactics are drawn from evasion strategies observed in real-world phishing campaigns and are detailed in Table 1.

After detection, messages undergo normalization to reconstruct the intended meaning. This step is crucial for restoring semantic clarity and preparing the message for downstream analysis.

**Table 1: Adversarial (Technical) Cues Used in Evasion Techniques**

| Cue | Description | Example |
|---|---|---|
| Character Spacing / Fragmentation | Words broken up or punctuated unnaturally. | "L o g i n", "U.r.g.e.n.t" |
| Homoglyphs | Similar-looking Unicode characters. | "ρ𝓖ʏpa𝓜", "□pp𝓔" |
| Misspellings | Strategic typos that preserve pronunciation. | "verif1cation", "acc0unt" |
| Symbol Insertion | Uses symbols or emojis to alter keywords. | "🔒 secure", "ex@mp3e" |
| Slang / Informal Tone | Uses casual speech and emojis to mimic peer behavior. | "Yo, wassup? 💬" |
| Char Swap / Noise Injection | Minor swaps or noise in phishing terms. | "clikc" for "click" |

Psychological manipulation is at the heart of smishing and other Social Engineering campaigns. Smishlock Holmes categorizes these manipulative elements as psychological cues, aligned with recognized social engineering techniques. Table 2 presents the six core cues used in this system, refined from phishing literature and grounded in behavioral security principles.

These cues are detected after normalization and used to interpret the manipulative framing of the message. Unlike evasion techniques, these cues target the human layer, exploiting familiarity, authority, and emotional pressure to provoke action.

**Table 2: Social Engineering / Psychological Manipulation Cues**

| Cue | Description | Example |
|---|---|---|
| Contextual Familiarity | Mentions real-world services or routines. | "Your Amazon package has shipped." |
| Impersonation | Claims identity of a known entity. | "This is Apple Support." |
| Urgency / Scarcity | Pressure to act within a limited time. | "Respond in 10 minutes!" |
| Authority Framing | Implies power from institutions or enforcement. | "Your case is with the IRS." |
| Personalization | Mentions user's name or specific details. | "Hi Andre, confirm now." |
| Contextual Language Consistency | Checks if tone fits the sender. | "Yo, your tax refund is ready" |

Beyond manipulation, effective smishing requires an execution mechanism. These are referred to as exploitation cues, representing clear paths for the attacker to engage the victim in malicious behavior—whether through redirection, data theft, or continued interaction.
These cues are essential in determining whether a message is actionable for attackers. Smishlock Holmes uses the presence of these cues to determine whether a message crosses into active threat territory

Table 3 outlines the recognized exploitation cues, including both direct and obfuscated forms. Notably, shortened or obfuscated URLs—which serve both evasion and exploitation purposes—are categorized here due to their execution potential.

**Table 3: Exploitation Cues**

| Cue | Description | Example |
|---|---|---|
| External Redirect | URL or link to an external site. | "Click here: http://bit.ly/login" |
| Continuation Request | Requests the user to reply or interact. | "Please text back to confirm." |
| Payment Request | Requests money, fees, or payment info. | "Your €57 fine is due." |
| Personal Data Request | Asks for ID, credentials, or sensitive info. | "Send your SSN to verify." |
| Obfuscated/Shortened URLs | Hides true link destination. | "http://tinyurl.com/bank-verify" |

Each of these cue types forms the foundation for Smishlock Holmes' multi-step classification and explanation system, ensuring that messages are not only labeled but also explained causally through a chain-of-cue interaction model. This layered approach reflects a hybrid of adversarial evasion detection, psychological manipulation awareness, and exploit path inference, making it suitable for both academic evaluation and user-facing threat explanation.

## 3.2 Explanation and Classification

Smishlock Holmes introduces a structured explanation mechanism to bridge the gap between classification outcomes and user comprehension. Unlike traditional binary classifiers, which merely flag messages as malicious or benign, this framework generates clear, context-aware explanations that articulate both the reasoning behind the classification and the cues that informed it.
The explanation process follows a fixed schema aligned with the reasoning steps used by the system:

1. **Technical cues detection** – identify obfuscation, formatting, or evasion patterns.
2. **Normalization** – rewrite obfuscated content for clarity.
3. **Scam cues detection** – identify psychological and exploitation indicators.
4. **Explanation** – link all detected cues together, describing how they interact to influence risk.
5. **Classification** – assign a verdict of *likely benign* or *suspicious* based on cue analysis.
6. **Meta-validation** – confirm that the explanation and classification align with the actual message.

For example, given the message:

*"fixed orthodontic workshop rd batch from diagnosis to tt planning hands on model and patient to may venue lucknowcontact for details"*

| Cue Type | Cue Name | Description |
|---|---|---|
| Psychological Cues | Contextual Familiarity | Mentions a plausible event/workshop, creating legitimacy within the medical/dental context. |
| Exploitation Cues | Continuation Request | Asks user to contact for details, potentially initiating dialogue without verifying sender legitimacy. |

**Explanation:**
"The message appears to promote a legitimate orthodontic workshop, with professional context and no technical or linguistic obfuscation. However, it ends with a vague 'contact for details' without identifying the sender or providing a verifiable source. This lack of sender verification, coupled with a prompt to engage, aligns with a suspicious classification as it could serve as a bait for further interaction."
This approach ensures that explanations remain factual, user-friendly, and actionable, allowing users to understand not just the classification result, but the precise reasoning that underpins it.

**Classification:** Suspicious

## 3.3 Dataset

The dataset used to evaluate Smishlock Holmes was derived from the Super SMS Dataset, a large real-world corpus containing both smishing and non-smishing messages. The original dataset is inherently imbalanced, with legitimate (non-smishing) messages forming the majority class, a distribution commonly observed in operational environments (Forman & Scholz, 2010).

To reduce the risk of performance metrics being biased toward the majority class, a balanced evaluation subset of 200 SMS messages was created by randomly sampling 100 smishing and 100 non-smishing messages from the corpus. This subset serves as the primary dataset for accuracy measurement, ensuring equal representation of both classes and enabling unbiased performance comparison.

### 3.4 Evaluation

The evaluation of Smishlock Holmes employs a dual-criteria approach combining a quantitative performance metric with qualitative explanation assessment.

Classification Accuracy: The primary quantitative metric is accuracy, calculated as the proportion of correctly classified messages in the balanced evaluation subset. This metric is chosen because the dataset contains equal numbers of smishing and non-smishing messages, ensuring that accuracy reflects performance across both classes without bias from class imbalance.

Explanation Quality (G-EVAL Framework): Qualitative assessment is performed using the G-EVAL framework, which scores each model-generated explanation along three dimensions:

- Correctness – Whether the explanation accurately reflects the cues, logic, and reasoning that justify the classification decision.
- Completeness – Whether the explanation addresses all relevant cues identified during analysis.
- Clarity – Whether the explanation is clear, concise, and understandable to a non-technical audience.

Each G-EVAL dimension is scored on a 0–5 scale. By combining classification accuracy with structured qualitative scoring, the evaluation assesses both what decision the system makes and how effectively it communicates the reasoning behind that decision.

# 4 Design Specification

Smishlock Holmes advances smishing detection from a binary, opaque classification to an AI–human-in-the-loop framework. It empowers end-users and analysts through explainability, highlighting the precise cues that make a message suspicious, thereby increasing trust in the detection pipeline and enabling informed decision-making.

The system architecture emphasizes modularity, regulatory compliance, and future-proof deployment, supporting both cloud-based and fully on-device inference.

## 4.1 Modularity

The modular architecture is designed around separation of concerns, enabling independent development, testing, and replacement of components. This approach ensures scalability, maintainability, and straightforward integration with third-party APIs. The main elements are:

- Embedded Chain-of-Thought (CoT) Reasoning

A novel prompt engineering technique that embeds reasoning steps directly into a fixed JSON schema. This improves classification robustness, enforces deterministic structure, and enables the generation of clear, user-friendly explanations that can be parsed, audited, and reused across different LLM backends.

- Output Enforcement

Strict server-side schema validation ensures that all LLM outputs are predictable, machine-readable, and safe for downstream processing. Any non-compliant output is flagged and rejected, preserving consistency and reliability.

- Model Agnosticism

The architecture supports seamless LLM backend swaps without modifying API logic. It also allows the addition or replacement of reasoning or auditing modules.

- Integration Readiness

The modular design exposes clear API hooks for integrating third-party analysis services, enrichment APIs, or enterprise-specific governance modules without altering the core detection pipeline.

## 4.2 Compliance with AI Regulations for Cybersecurity Applications

In accordance with the EU Artificial Intelligence Act (AI Act) and guided by the Seven Principles of Trustworthy AI, Smishlock Holmes aligns with the following:

- Human Agency and Oversight – Supports analyst review and override of automated classifications, ensuring human decision-making authority is preserved (GDPR Recital 71, Art 22).

- Robustness and Safety – Employs schema validation, input sanitisation, and drift detection to ensure reliable, resilient, and safe operation under adversarial or unexpected inputs (GDPR Art 22).
- Privacy and Data Governance – Processes only message text, excludes metadata, and uses ephemeral storage to protect personal data, giving users control over their information (GDPR Art 13, 14, 15).
- Transparency – Produces a detailed JSON reasoning trace for each inference, enabling traceability and explanation of decisions (GDPR Recital 71).
- Diversity and Fairness – Designed for generalised performance across varied message types and scam scenarios; does not rely on attributes tied to specific demographics (GDPR Art 22).
- Societal and Environmental Well-being – Contributes to digital security and user empowerment against fraud, promoting societal resilience (GDPR Art 13, 14, 15).
- Accountability – Maintains auditable logs, clear processing logic, and consistent schema enforcement, ensuring that outputs can be reviewed and challenged (GDPR Art 13, 14, 15).

# 5 Implementation

The Smishlock Holmes framework requires a large language model capable of classifying smishing messages and generating layered, interpretable explanations in natural language. GPT-4.1 was selected for the prototype stage primarily due to its ease of deployment, robust instruction-following, and proven reliability in structured reasoning tasks. Since the innovation in this work lies in the prompt design and explainable framework rather than the base model itself, GPT-4.1 enabled rapid prototyping and evaluation without infrastructure friction. Quantized open-source models will be tested in future work to explore on-device inference.

## 5.1 Software Architecture

The system is modular and designed for real-time, explainable classification:

- Programming Language: Python 3.12
- LLM Orchestration: LangChain for prompt handling and chain logic.
- Model Backend: GPT-4.1 via OpenAI API.
- Interface Layer: FastAPI for the /analyze and /geval REST endpoints.
- Prompt Engine: Implements a JSON-Embedded Few-Shot CoT format.
- Post-Processing: Validates schema.
- Logging: All detection outputs and evaluation results are stored in a JSONL log file, preserving the original message, model output, and evaluation scores for reproducibility, audits, and performance analysis.
-

The architecture allows:

- Swapping LLMs without changing API logic.
- Adding reasoning modules (e.g., URL or image analysis).
- Integrating external threat intelligence APIs.
- Extending evaluation criteria by modifying the G-Eval JSON schema.

This architecture enables Smishlock Holmes to evolve alongside advances in LLMs and secure AI infrastructure.

## 5.2 Prompt Architecture

The prompt architecture in Smishlock Holmes is designed to operationalise the layered reasoning methodology described in Sections 3.1 and 3.2 while ensuring deterministic, machine-readable outputs suitable for integration into automated security workflows. This architecture is built around a JSON-Embedded Few-Shot Chain-of-Thought (CoT) format, enabling the large language model (LLM) to reason in discrete, traceable steps and output both classification results and explanations within a fixed schema.

The core design principle is schema-constrained reasoning: each reasoning step is explicitly defined in the prompt and bound to a structured output field. This enforces alignment between the model's intermediate reasoning and its final classification. The prompt enforces six sequential stages corresponding to the system's analytical pipeline:

1. Technical Cue Detection – Identify adversarial evasion tactics (e.g., homoglyphs, spacing, symbol insertion) as per the taxonomy in Table 1.
2. Normalization – Restore semantic clarity of obfuscated text.
3. Scam Cue Detection – Detect psychological manipulation and exploitation cues from Tables 2 and 3.
4. Explanation Generation – Provide a causal narrative linking detected cues to the threat likelihood.
5. Classification – Assign a binary verdict (likely benign / suspicious) based on cue analysis.
6. Meta-Validation – Cross-check that the explanation logically supports the classification.

The few-shot component of the prompt supplies representative examples for each reasoning step, selected from real smishing cases and benign messages.

To enhance robustness against prompt injection or output drift, the architecture incorporates output enforcement hooks in the post-processing layer (Section 5.1). Any model response that does not conform to the JSON schema is automatically rejected and re-queried. This mechanism ensures compliance with both robustness and accountability principles under the EU AI Act, as every decision is traceable to a structured reasoning record.

# 6 Evaluation

This section presents the evaluation of Smishlock Holmes against the dataset described in Section 4, assessing classification performance, explanation quality, and the role of meta-validation. The goal was to determine how effectively the system detects smishing messages and communicates its reasoning in a manner suited to cognitively constrained mobile contexts.

Across 200 SMS samples (balanced between suspicious and likely benign), Smishlock Holmes achieved an **overall accuracy of 87.5%**, with:

- **20 false negatives** – largely clean-formatted, brand-associated messages with subtle malicious intent.
- **5 false positives** – mostly casual or chain-style benign messages flagged as suspicious.

Misclassifications followed distinct patterns:

- **False negatives** arose from narrow exploitation cue definitions that overlooked low-signal but high-risk features (e.g., brand impersonation without links).

- **False positives** resulted from over-weighting weak cues, such as informal tone or generic continuation requests.

Explanations were evaluated using the G-Eval framework (0–5 scale):

- **Correctness**: 4.68 – Strong alignment between identified cues and the associated risks.
- **Completeness**: 3.94 – Most variability; secondary cues and context-dependent risks were sometimes omitted.
- **Clarity**: 4.50 – Readable and logically structured, though occasionally generic in emphasis.

These results suggest that explanations are both accurate and accessible but could be more informative by addressing a broader range of relevant cues.

## 6.1 Discussion

The evaluation confirms that Smishlock Holmes delivers robust detection and clear, trustworthy explanations in a mobile context, aligning with the project's research aim. However, two limitations emerged:

1. Cue Definition Narrowness – A narrow exploitation cue taxonomy, causing increased false negatives due to implicitly exploitable paths.
2. Meta-Validation Ineffectiveness – The meta-validation stage, instead of serving as a challenge mechanism, merely confirmed the model's initial output.

These issues are not inherent flaws of the LLM-based approach but rather refinements in configuration and contextual integration. Addressing them would require:

- Expanding cue taxonomies to incorporate nuanced indicators.
- Redesigning meta-validation as a multi-agent process.

When placed in the context of Section 2.4, these findings illustrate that Smishlock Holmes embodies many of the LLM advantages in security detection—generalisation to obfuscation, combined classification and explanation—but also inherits known challenges around prompt sensitivity, contextual trust calibration, and explanation completeness. Refining these areas will strengthen both detection reliability and user trust, fulfilling the system's objective of delivering explainable, mobile-optimised smishing defence.

# 7 Conclusion and Future Work

This research set out to answer the question: How can Large Language Models (LLMs) be applied to detect smishing messages in a way that not only improves classification robustness but also generates interpretable explanations that enhance user awareness and trust, particularly in cognitively constrained mobile contexts?

To address this, Smishlock Holmes was developed as a modular, explainable AI framework that integrates adversarial robustness and user-friendly narrative explanations into a unified detection pipeline. The prototype leveraged GPT-4.1 within a JSON-Embedded Few-Shot Chain-of-Thought architecture, enabling classification and explanation in a structured, auditable format. Evaluation combined a balanced gold-label dataset for accuracy testing and the G-Eval framework for explanation quality assessment.

The system achieved 87.5% classification accuracy, with high scores for correctness (4.68/5) and clarity (4.50/5) of explanations, confirming its capacity to deliver trustworthy, context-aware outputs. However, completeness (3.94/5) revealed room for improvement.

Key limitations included: (1) gray-area messages, texts that cannot be definitively labeled as benign or malicious without access to sender identity or additional context (2) a meta-validation stage that reinforced, rather than critically challenged, initial outputs (3) the

reliance on cloud-based LLMs raises privacy and regulatory considerations, particularly under GDPR.
Implications of this research extend to both academic and operational domains. Academically, it demonstrates that combining classification and explanation within a single LLM backbone is feasible and effective for mobile-centric security contexts. Operationally, the modular design supports GDPR-aligned, explainable defences that can integrate into messaging platforms, enterprise tools, or end-user applications.

**Future Work:**

- Cue Taxonomy Expansion – Broaden exploitation cue definitions to capture multi-message scam patterns.
- Contextual Metadata Integration – Incorporate sender identity, historical interaction data, and threat intelligence to improve risk assessment.
- Self-Auditable Multi-Agent Verification – Replace meta-validation with a multi-agent process that enables independent critique and revision of outputs using criteria derived from frameworks such as G-Eval.
- On-Device Inference – Transition to quantized open-source models for privacy-preserving offline use.
- User Experience Studies – Conduct trials to measure the real-world impact of explanations on detection behaviour and trust calibration.
- Commercialisation Pathways – Explore SaaS or API-based delivery for integration with carriers, messaging platforms, and SOC tools.

By pursuing these directions, Smishlock Holmes can evolve from an integration-ready endpoint into a scalable, privacy-conscious, self-auditable, and user-adaptive smishing defence system for emerging mobile communication threats.

# References


Atf, Z. and Lewis, P.R., 2025. Is trust correlated with explainability in AI? A meta-analysis. *IEEE Transactions on Technology and Society*. https://doi.org/10.1109/TTS.2025.3558448

Baron, S., 2025. Trust, explainability and AI. *Philosophy & Technology*, 38(4). https://doi.org/10.1007/s13347-024-00837-6

Chamola, V., Hassija, V., Sulthana A, R., Ghosh, D., Dhingra, D. and Sikdar, B., 2023. A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3294569

Ferrario, A. and Loi, M., 2022. How explainability contributes to trust in AI. In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 21–24 June 2022, Seoul, Republic of Korea. ACM, New York, NY, USA. https://doi.org/10.1145/3531146.3533202

Figl, K. and Remus, U., 2023. Thinking fast and thinking slow: Digital devices' effects on cognitive reflection. *Journal of Management Information Systems*, 40(2), pp.580–623. https://doi.org/10.1080/07421222.2023.2196769

Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. and Park, P.S., 2024. Devising and detecting phishing emails using large language models. *IEEE Access*, 12, pp.42131–42144. https://doi.org/10.1109/ACCESS.2024.3375882

Labarta, T., Kulicheva, E., Froelian, R., Geißler, C., Melman, X. and von Klitzing, J., 2024. Study on the helpfulness of explainable artificial intelligence (XAI). In: L. Longo, S. Lapuschkin and C. Seifert, eds. *Explainable Artificial Intelligence*. Communications in Computer and Information Science, vol. 2156. Springer, Cham. https://doi.org/10.1007/978-3-031-63803-9_16

Liao, M., Wang, J., Chen, C. and Sundar, S.S., 2025. Less vigilant in the mobile era? A comparison of information processing on mobile phones and personal computers. *New Media & Society*, 27(5), pp.2657–2683. https://doi.org/10.1177/14614448231209475

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. and Zhu, C., 2023. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*. Association for Computational Linguistics, Singapore. https://doi.org/10.18653/v1/2023.emnlp-main.153

Lyu, Q., Havaldar, S., Stein, A., Zhang, L., Rao, D., Wong, E., Apidianaki, M. and Callison-Burch, C., 2023. Faithful Chain-of-Thought Reasoning. In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Asia-Pacific Chapter of the Association for Computational Linguistics*, Nusa Dua, Bali. pp. 305–329. https://doi.org/10.18653/v1/2023.ijcnlp-main.20

Papenmeier, A., Kern, D., Englebienne, G. and Seifert, C., 2022. It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction*, 29(4), Article 35. https://doi.org/10.1145/3495013

Proofpoint, 2024. *State of the phish 2024*. [online] Available at: https://www.proofpoint.com/us/resources/threat-reports/state-of-phish [Accessed 11 August 2025].

Rahman, M.L., Timko, D., Wali, H. and Neupane, A., 2023. Users really do respond to smishing. In: *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy (CODASPY '23)*, 24–26 April 2023, Charlotte, NC, USA. ACM. https://doi.org/10.1145/3577923.3583640

Rashid, F., Ranaweera, N., Doyle, B. and Seneviratne, S., 2025. LLMs are one-shot URL classifiers and explainers. *Computer Networks*, 258, p.111004. https://doi.org/10.1016/j.comnet.2024.111004

Salman, M., Ikram, M., Basta, N. and Kaafar, M.A., 2025. SpaLLM-Guard: Pairing SMS spam detection using open-source and commercial LLMs. *arXiv preprint* arXiv:2501.04985.

Shin, J. and Chan-Olmsted, S., 2023. User perceptions and trust of explainable machine learning fake news detectors. *International Journal of Communication*, 17, pp.518–540.

Tabassum, S., Faklaris, C. and Lipford, H.R., 2024. What drives SMiShing susceptibility? A U.S. interview study of how and why mobile phone users judge text messages to be real or fake. In: *USENIX Symposium on Usable Privacy and Security (SOUPS) 2024*, 11–13 August 2024, Philadelphia, PA, USA.

Timko, D., Hernandez Castillo, D. and Rahman, M.L., 2025. Understanding influences on SMS phishing detection: User behavior, demographics, and message attributes. *Symposium on Usable Security and Privacy (USEC) 2025*, San Diego, CA. https://dx.doi.org/10.14722/usec.2025.23027