Generating Threat Modeling Antipersonas with Large Language Models: A Feasibility Study

Anna Cristina Ferreira Alves
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brazil
annacfa@ic.ufrj.br

Rafael Maiani de Mello
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro, Brazil
rafaelmello@ic.ufrj.br

Abstract

Early identification of security threats during software requirements engineering is a critical challenge, often hindered by limited security expertise and time constraints. This work investigates the ability of LLMs to automatically generate antipersonas from software requirements. We propose a structured antipersona representation and a prompt-based approach grounded in STRIDE and MITRE ATT&CK frameworks, to guide LLMs in generating antipersonas, and we conducted a preliminary feasibility study to evaluate the structural completeness, specificity, and adherence to these frameworks of the generated antipersonas. The generated antipersonas were predominantly assessed as structurally complete, specific, and largely adherent to STRIDE and MITRE ATT&CK, though adherence to STRIDE showed low inter-rater agreement, and we also identified recurring limitations in framework adherence. The findings suggest that LLMs are feasible for generating well-structured antipersonas that can support, as a preliminary step, early threat identification during requirements engineering, though careful human validation against real-world threats remains necessary to ensure the practical relevance of the generated antipersonas.

Keywords

Threat Modeling, Antipersona, Large Language Models, Prompt Engineering, Security Requirements

1 Introduction

In recent years, the use of LLMs has expanded significantly, driven by their potential to automate tasks that require a high degree of knowledge and specialization. The application of these models in the field of cybersecurity has been increasingly explored, particularly in tasks such as the generation of attack scenarios and attack paths [11]. Such investigations are not limited to identifying possible applications; they also seek to understand the limitations of these models, including their tendency to produce generic responses and their lack of reliability in critical contexts [6].

Unlike tasks where standardized responses are sufficient, the usefulness of security artifacts depends directly on their specificity. Vague representations of malicious agents, such as *"a financially motivated attacker with moderate technical capability"*, do not provide sufficient detail to distinguish between different attacker profiles, thereby limiting their value for threat modeling and security decision-making. Consequently, investigating approaches that enable LLMs to generate more specific and actionable descriptions of malicious agents is an important research objective for advancing the effective application of Generative AI in cybersecurity.

The representation of malicious agents can be addressed through the concept of *antipersonas*. Unlike attack trees or misuse cases, antipersonas maintain a persistent representation of an adversary across multiple requirements, allowing analysts to reason consistently about attacker motivations, capabilities, and evolving attack strategies. Although the concept of *persona* is well established and widely adopted in the field[9], the concept of *antipersona* still lacks a single, widely accepted definition [3, 8, 14]. This work defines an antipersona as an integrated, framework-aligned representation of an attacker's capabilities, tactics, and motivations for exploiting software system vulnerabilities. This definition emphasizes actionable characteristics that can support systematic threat modeling and security analysis while providing a structured target for LLM-based generation.

This work introduces a prompt-based approach for generating antipersonas and reports a preliminary study conducted to evaluate its feasibility through the qualitative analysis of antipersonas generated in the context of a real-world open-source software system. To this end, quality is operationalized across three verifiable dimensions: (i) structural completeness, (ii) adherence to STRIDE and MITRE ATT&CK frameworks; and (iii) specificity. We applied this approach to forty antipersonas generated across four independent executions targeting a real-world system. The general results revealed predominantly satisfactory completeness and specificity and largely satisfactory framework adherence alongside recurring semantic inconsistencies.

2 Background and Related Work

This section briefly introduces the key concepts underlying this research, including antipersonas and the STRIDE and MITRE ATT&CK frameworks, and positions the present work in relation to prior work on the generation of abuser stories.

2.1 Antipersona

The term *anti-persona* was earlier coined by Steele and Jia [14] within their proposal of Adversary Centered Design, in which anti-scenarios, anti-use cases and anti-personas were introduced as duals of standard user-centered design artifacts for modeling threats to electronic voting systems. That formulation, however, remains lightly structured and does not integrate attackers' capabilities, tactics and motivations. Furthermore, it does not address security framework coverage (e.g., STRIDE and MITRE ATT&CK) into a single, verifiable representation intended to support systematic threat-modeling decisions – which is the gap this work addresses.

Related concepts such as threat actor profiles [8] and adversary personas used in red team exercises [3] cover partial aspects of this space: the former focus on observable attributes of real actors,

while the latter involve situational simulations of behavior. However, neither of these concepts, in isolation, provides a structured representation that integrates capabilities, tactics, and motivations of malicious agents with the explicit purpose of supporting threat modeling activities.

Unlike adjacent concepts such as threat actor profiles and adversary personas – including the original notion of *anti-persona* introduced by Steele and Jia [14] – an antipersona as defined here integrates, within a single representation, the agent's capabilities, tactics, and motivations, with the explicit purpose of guiding the identification and prioritization of threats in systems, websites, or applications. None of these individual elements is novel in isolation: data-grounded attacker personas already exist [3], and so does LLM-based threat modeling [7, 11]. The contribution of this work lies instead in the synthesis, that is, in integrating capability, motivation, and framework-verifiable fields (STRIDE and MITRE ATT&CK) into a single, structured artifact – a combination that, to our knowledge, has not been jointly operationalized in prior work.

2.2 STRIDE

STRIDE is a security threat classification model developed by Microsoft to support threat modeling activities [12]. The model is based on a central premise: to protect a system, it is necessary to understand the ways in which it can be attacked [12]. In this way, STRIDE works as a tool for structuring adversarial thinking, guiding analysts to systematically examine each component of a system through six threat categories: spoofing, tampering, repudiation, information disclosure, denial of service, and elevation of privilege, each associated with the violation of a classical security principle.

Spoofing refers to identity forgery, such as when an attacker impersonates a legitimate user to gain unauthorized access. *Tampering* concerns the unauthorized modification of data, such as altering records in a database. *Repudiation* involves the denial of performed actions, such as a user claiming that they did not execute a transaction. *Information Disclosure* corresponds to unauthorized access to confidential information, such as intercepting data in transit. *Denial of Service* refers to attacks that render a system unavailable to its legitimate users. Finally, *Elevation of Privilege* occurs when a malicious agent obtains permissions beyond those originally granted, compromising the system's access control.

In the practice of threat modeling, STRIDE is typically applied in conjunction with data flow diagrams, where each system element, such as processes, data stores, and data flows, is analyzed in light of the six categories. This process enables security teams to comprehensively identify threats and prioritize controls even before system implementation [12].

2.3 MITRE ATT&CK

MITRE ATT&CK is a widely adopted cybersecurity knowledge framework that describes, organizes, and analyzes the real-world behavior of adversaries throughout the lifecycle of an attack [15]. ATT&CK, which stands for Adversarial Tactics, Techniques, and Common Knowledge, provides a structured repository of tactics and techniques derived from real-world incidents. Rather than focusing on vulnerabilities, the framework characterizes adversary behavior by specifying what attackers do, how they do it, at which stage of an attack, and the known variations of these actions [15].

In the context of threat modeling, MITRE ATT&CK provides a standardized vocabulary for representing threats in a precise and verifiable manner. By mapping an adversary's behavior to ATT&CK tactics and techniques, analysts can compare threat representations, identify gaps in security control coverage, and communicate risks consistently across stakeholders.

2.4 Related Work

Recent studies have increasingly explored the use of Large Language Models (LLMs) to support early security analysis from software engineering artifacts. While most research on LLM-based security focuses on source code analysis, recent efforts have shifted security activities to earlier phases of the software lifecycle. For instance, [7] proposed an LLM-assisted threat modeling approach that analyzes architectural design documents together with National Vulnerability Database (NVD) data to support security reasoning during the design phase, demonstrating that Retrieval-Augmented Generation (RAG) improves response quality while reducing hallucinations. Similarly, [4] integrated frontier LLMs into Security Chaos Engineering to automatically derive attack-defense trees from user stories, showing that LLMs can leverage cybersecurity knowledge embedded during pre-training, although expert validation remains necessary. More closely related to our work, [2] investigated the automatic generation of abuser stories and misuse cases from software requirements using different LLMs under zero-shot and one-shot prompting strategies. Their evaluation, based on 907 generated abuser stories validated against an expert-crafted oracle grounded in the OWASP ASVS standard, showed that LLMs reliably identify explicit, requirement-level threats but consistently miss threats that depend on implicit architectural knowledge.

Our study differs substantially from these previous works in both objective and artifact. Rather than generating threat-oriented artifacts from requirements or design documents, we investigate the automated generation of antipersonas, which constitute persistent and structured representations of malicious actors. In our approach, antipersona integrates identity, motivations, capabilities, behaviors, and attack vectors while maintaining semantic consistency across multiple security frameworks, including STRIDE and MITRE ATT&CK. Consequently, the task extends beyond identifying isolated threats to constructing coherent adversarial profiles that remain distinguishable from one another throughout the system. This broader representational scope introduces challenges that are not addressed by prior studies on threat modeling, attack-defense trees, or abuser story generation.

3 Antipersona Structure

To enable the systematic generation and evaluation of antipersonas, it is necessary to define a standardized structure that consistently represents the relevant attributes of a malicious agent. This structure must be sufficiently expressive to capture adversarial motivations, capabilities, and behaviors while supporting systematic threat modeling.

Table 1 presents the proposed antipersona structure. The template defines the fields that compose a complete antipersona, organized into seven analytical dimensions: *identity*, *motivations*, *capabilities*, *adversarial behavior*, *attack vectors and target assets*, *framework adherence*, and *impact and relevance*. Each dimension groups conceptually related attributes that describe complementary aspects of a malicious agent, providing a consistent representation suitable for both automated generation and qualitative evaluation.

Table 1: Dimensions of the Proposed Antipersona Model

Dimension	Fields	Description
Identity	id, name, summary, type	Defines the core identification attributes of the antipersona
Motivations	primary motivations, secondary motivations, objectives	Captures adversarial intent and goals
Capabilities	technical level, resources, access level, knowledge	Represents the attacker's ability to execute threats
Adversarial Behavior	persistence, stealth, risk tolerance, observed patterns	Describes behavioral traits relevant to threat modeling
Vectors and Assets	attack vectors, target assets	Defines how and where attacks are performed
Framework Adherence	mitre att&ck, stride coverage	Ensures alignment with established frameworks
Impact and Relevance	potential impacts, threat modeling relevance, specificity markers	Captures the expected consequences of attacks, their priority for threat modeling, and attributes that distinguish attacker profiles.

To operationalize the proposed structure, we developed a JSON-based template. JSON was selected because it provides a hierarchical representation that is both machine-readable and easily interpretable by humans, facilitating its integration into prompt-based workflows and the automated processing of the generated artifacts.

Beyond standardizing the representation of malicious agents, the proposed template introduces explicit classifications that support both generation and evaluation. It organizes attacker characteristics into well-defined analytical dimensions and incorporates categorical attributes, including attacker type, technical capability, persistence, stealth, risk tolerance, and threat priority. Furthermore, it explicitly associates adversarial behavior with the STRIDE and MITRE ATT&CK frameworks while capturing potential impacts, recommended security controls, and specificity markers. Together, these elements encourage the generation of differentiated and actionable antipersonas rather than generic attacker descriptions, while providing objective criteria for assessing the structural completeness, framework adherence, and specificity of the generated artifacts.

4 Study Design

The present feasibility study aims to *evaluate the feasibility of the proposed LLM-based approach in generating structurally complete, specific, and framework-adherent antipersonas*. The study follows a structured workflow composed of three main stages: (1) the use of a widely documented system as input; (2) the application of prompt engineering techniques; and (3) the qualitative evaluation of the generated antipersonas based on established information security frameworks. The study workflow is illustrated in the figure below.

4.1 Documented System

OpenMRS is a global open-source Electronic Health Record system widely used in low-income healthcare settings around the world. It was selected for this study because it brings together three methodologically relevant characteristics: rich and publicly accessible technical documentation, which enables its use as a controlled input in the experiment; an open-source nature, which ensures the reproducibility of the study; and a sufficiently varied attack surface to allow for the generation of distinct and differentiated antipersonas.

The data source used in this study is the official wiki of the public technical documentation of the OpenMRS system.[1] To ensure a controlled and reproducible input, the sections relevant to the identification of potential attackers were filtered, including Core Architecture (EAV Database, Concept Dictionary), Context and APIs (REST and FHIR), Authorization and Authentication, Modules, and Frontend Tech Stack. This context addresses the components with the greatest exposed attack surface — encompassing clinical data, public endpoints, access control, and system extensibility.

4.2 Large Language Model

The LLM used in this study is Claude Sonnet 4.6, developed by Anthropic and released on February 17, 2026 [1]. The model represents a full update of the intermediate tier of the Claude family, with improvements in long-context reasoning, agentic planning, and a context window of one million tokens. This capacity enables the processing of extensive technical documentation in a single inference, without input fragmentation — a relevant condition for this study, which uses OpenMRS documentation as controlled input. The model also introduces an adaptive thinking mode, in which developers can control the level of reasoning effort through a configurable parameter, allowing the model to allocate varying degrees of processing according to task complexity [1].

The choice of Claude Sonnet 4.6 over alternative models is further justified by its balance between performance and suitability for the information security context. Tests documented in the model's system card indicate a refusal rate of 99.38% for explicitly malicious requests, while a success rate of 91.78% was recorded for benign dual-use technical requests — indicating that the model is capable of responding to legitimate threat modeling tasks without the adversarial nature of the representation triggering undue refusals [1]. Additionally, the model was deployed under Anthropic's AI Safety Level 3 (ASL-3) standard and demonstrated significant improvement in resilience to prompt injection attacks relative to its predecessor, achieving performance equivalent to that of Claude

[1] https://openmrs.atlassian.net/wiki/

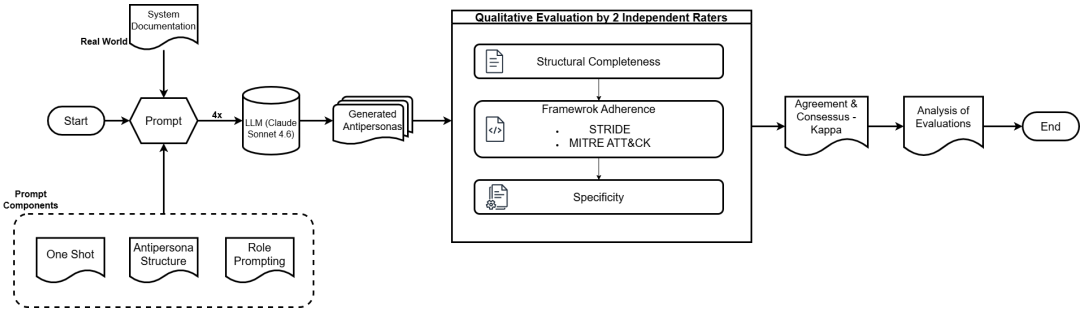

Figure 1: Overview of the study workflow.

Opus 4.6 in this dimension [1]. Compared to Opus 4.6, Sonnet 4.6 represents the optimal trade-off between capability, cost, and suitability for the task of generating structured JSON artifacts from technical documentation [1].

4.3 Prompt Engineering

Prompt engineering consists of the technique of crafting instructions for language models with the objective of obtaining more precise, useful, and goal-aligned responses [10]. In this study, two complementary techniques were combined. The first is one-shot prompting, which consists of providing the model with a single example of the expected output [5]. This methodological choice was grounded in the premise that the JSON schema already establishes the structural rigor required for the response, making the demonstration of a specific narrative level sufficient for the objectives of the study. Consequently, the application of few-shot prompting — which involves presenting multiple examples — was discarded as redundant. The inclusion of several examples could, in fact, limit the diversity of the generated antipersonas by inducing an undesirable convergence in the outputs. Similarly, the chain-of-thought technique, which makes the model's reasoning process explicit [16], was not employed, since the priority was to obtain a purely structured output in JSON format.

The second technique applied was the assignment of an epistemic role to the model (role prompting), positioning it as an expert analyst in information security and threat modeling, with the objective of directing its knowledge toward the relevant domain. The two techniques operate in a complementary manner: the role defines the knowledge profile mobilized by the model, while the one-shot example anchors the format and depth of the expected output. The complete prompt is available in the supplementary material.

4.4 Experimental Procedure

The experiment was conducted using the Web interface of Claude Sonnet 4.6. The extended thinking mode was enabled with the objective of prompting the model to generate more specific and well-grounded representations, given that the task requires non-trivial adversarial inference. Web search was disabled to ensure that the model relied solely on the content available in the prompt, without influence from external sources. Each generation was performed in a single interaction, without subsequent refinement, ensuring that all antipersonas were produced under identical conditions.

The prompt instructed the model to generate ten distinct antipersonas, a fixed number defined to favor broad coverage of the system's attack surface without compromising the distinction between the generated profiles. Fixing n within the prompt itself eliminates variability in the number of outputs across executions, thereby reinforcing reproducibility.

Four independent executions were performed using the same prompt and identical experimental conditions. Each execution was started in a new conversation to avoid contextual carry-over effects. The objective was to evaluate the stability of the generation process across repeated interactions. Repeated executions were intended to increase the diversity of the generated antipersonas and reduce the risk that the conclusions would be based on a single favorable response.

4.5 Data Classification

The evaluation of the generated antipersonas was conducted qualitatively through human assessment (two researchers), structured around three dimensions: *structural completeness*, *framework adherence*, and *specificity*. Each dimension was evaluated using the same three-level scale: *satisfactory*, *partial*, and *unsatisfactory*.

Structural completeness was assessed by verifying adherence to the defined JSON template, requiring that all mandatory fields be fully populated, without placeholders, missing information, or generic values. Beyond field completeness, this criterion also required internal coherence among attributes, such that motivations, capabilities, behaviors, and attack vectors formed a consistent representation of a single agent. An antipersona was classified as *satisfactory* only when all of these conditions were met.

Specificity was assessed according to three operational criteria: (i) whether the summary alone uniquely identified the attacker profile; (ii) whether the generated profile could be clearly distinguished from the remaining antipersonas produced in the same execution; and (iii) whether the profile would plausibly lead to different security controls or mitigation priorities than those suggested by the other antipersonas. An antipersona was classified as *satisfactory* only when all three conditions were satisfied.

Framework adherence was evaluated through two independent sub-criteria. The STRIDE sub-criterion assessed the *stride_coverage* field according to two complementary dimensions: categorical validity and behavioral pertinence. Categorical validity required that all assigned categories belonged to the official STRIDE taxonomy

(Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege), while behavioral pertinence required that each assigned category be explicitly supported by the adversarial behaviors, attack vectors, and observed patterns described in the antipersona. Any valid STRIDE category lacking behavioral evidence was treated as an inconsistency, following the validation criteria presented in Table 2.

Table 2: STRIDE-based validation criteria

Level	Criterion
Satisfactory	All categories are valid and directly supported by the described behavior
Partial	Categories are valid, but one or more lack clear support in the modeled behavior
Unsatisfactory	Categories are invalid, absent, or unrelated to the described behavior

The MITRE ATT&CK sub-criterion assessed the *tactics_and_techniques* field across two dimensions: formal validity and contextual coherence. Formal validity required that the referenced tactic (*tactic_id*) and technique (*technique_id*) identifiers existed in the framework and matched their declared names. Contextual coherence evaluated whether the selected tactics and techniques were consistent with the antipersona profile, attack vectors, and behavioral patterns. Formally valid but overly generic techniques that lacked specificity for the modeled agent were considered partially adequate. Table 3 summarizes the criteria applied.

Table 3: MITRE ATT&CK-based validation criteria

Level	Criterion
Satisfactory	Valid IDs and techniques coherent with the declared attack vectors and behaviors
Partial	Valid IDs, but one or more techniques are generic or weakly justified by the profile
Unsatisfactory	Invalid IDs or techniques incompatible with the modeled behavior and context

4.5.1 Human Validation Procedure. Two evaluators, both with academic and practical backgrounds in cybersecurity and threat modeling, assessed all generated antipersonas according to the criteria described in the previous subsections. The evaluations were performed independently, without communication between evaluators. To assess the reliability of the qualitative evaluation, inter-rater agreement was measured using Cohen's Kappa coefficient [13]. Agreement was computed independently for each evaluation dimension before the consensus phase. Disagreements were resolved only after the independent assessments had been completed, which produced the final classification used in the analysis.

4.6 Threats to Validity

Although the evaluation protocol was designed to promote consistency, some limitations remain. First, the assessment relies on the judgment of the two evaluators, which inevitably introduces subjectivity – including possible differences in interpretation and prior expectations – despite the use of explicit evaluation criteria and independent assessment. Second, the study evaluates antipersonas generated from a single software system, which may limit the generalizability of the findings to other systems. Finally, the study was conducted using a single LLM configuration, and the results should be interpreted as preliminary evidence of feasibility rather than comparative superiority over alternative models.

5 Results

Before resolving disagreements through consensus, inter-rater agreement was assessed using Cohen's Kappa (Table 4). Due to the extreme imbalance in category distributions, with most evaluations classified as *satisfactory*, the dimensions of structural completeness and specificity yielded a Kappa value of zero despite the high observed agreement between raters.

The evaluation of STRIDE resulted in insufficient inter-rater agreement (40% observed agreement, $\kappa = -0.05$), indicating that, despite the predominance of *satisfactory* classifications reported below, raters frequently disagreed on individual cases. This outcome was primarily explained by divergent interpretations of the *Repudiation* threat category, which affected 35% of the generated personas. In contrast, the MITRE ATT&CK dimension achieved a statistically significant moderate level of inter-rater agreement.

Table 4: Inter-rater agreement for the evaluated dimensions.

Dimension	Obs. agreement	Cohen's κ	p-value
Completeness	97.5%	0.000*	–
Specificity	90.0%	0.000*	–
STRIDE	40.0%	−0.050	0.50
MITRE ATT&CK	82.5%	0.615	< 0.001

*The Kappa statistic is not meaningful due to the absence of variability in one evaluator's ratings (all instances were assigned to the same category). In these cases, the observed agreement should be interpreted instead of Cohen's Kappa.

5.1 Structural Completeness and Specificity

Regarding *structural completeness*, of the 40 antipersonas, 39 were classified as *satisfactory* and one as *partial*. All mandatory fields were completed, with no missing values, placeholders, or structural omissions. The generated artifacts exhibited consistent internal coherence, with motivations, capabilities, behaviors, and attack vectors forming cohesive and complete representations. These results indicate that combining a JSON schema with a populated example (one-shot) was sufficient for the model to consistently satisfy the structural requirements of the template. The high level of consistency across independent executions further suggests that this prompting strategy provides a stable approach for generating structured threat-modeling artifacts.

Regarding *specificity*, 36 of the 40 antipersonas received a *satisfactory* classification, and four were classified as *partial*, with no *unsatisfactory* cases. Across most executions, the summaries were sufficiently descriptive to identify the modeled agent without consulting the remaining fields, clearly conveying who the agent is,

what motivates them, and how they operate. Furthermore, no antipersona was considered indistinguishable from the others within the same execution, ensuring that each representation would support security control decisions distinct from those prompted by the other profiles in the generated set.

5.2 Framework Adherence

5.2.1 STRIDE. STRIDE-based validation assessed whether the categories listed in the *stride_coverage* field of each antipersona were formally valid and supported by the modeled behavior. Of the 40 antipersonas, 30 received a *satisfactory* classification, 9 were classified as *partial*, and one *unsatisfactory*.

In some *partial* cases, the model associated stealthy behavior, the use of legitimate accounts, or attempts to avoid detection with *repudiation*, although these behaviors reflected operational concealment rather than the ability to deny or prevent proof of an action. The second most frequent pattern involved *elevation of privilege*, which was assigned to adversaries who already possessed sufficient privileges, such as insiders or authenticated administrators, without any actual privilege escalation occurring. Less frequent inconsistencies involved spoofing, typically due to confusion between unauthorized access and identity impersonation. Another isolated partial case involved the omission of *information disclosure* from a ransomware profile whose objectives explicitly described pre-encryption data exfiltration. In the single *unsatisfactory* case, the *stride_coverage* field included *integrity*, a security principle related to *tampering* rather than a valid STRIDE category.

Overall, the findings indicate that the LLM tends to overestimate categories associated with stealth and privileged access, prioritizing the operational context of the attacker over the precise semantic boundaries defined by STRIDE. This suggests that more explicit guidance distinguishing attacker behaviors, security objectives, and STRIDE threat categories could improve classification accuracy.

5.2.2 MITRE ATT&CK. MITRE ATT&CK-based validation assessed the semantic consistency between the assigned techniques and the adversarial behaviors described in each antipersona. Of the 40 antipersonas, 24 received a *satisfactory* classification, and 16 were classified as *partial*, with no *unsatisfactory* cases. The most frequent errors involved semantically plausible but behaviorally unsupported technique assignments, indicating that the model often relied on conceptual similarity rather than the operational definitions of ATT&CK techniques. Recurring issues included confusion among related tactics (e.g., Discovery, Reconnaissance, and Collection) and the misuse of broad techniques such as *Valid Accounts*, *Exfiltration Over Web Service*, and *Web Shell*. These findings suggest that, although the generated antipersonas captured relevant adversarial behaviors, additional validation is required to ensure semantically accurate ATT&CK mappings, particularly at the subtechnique level. This echoes prior findings on LLM-generated security artifacts, where models reliably capture surface-level patterns but struggle with deeper contextual precision [2, 4].

6 Conclusion and Future Work

The study reported in this paper provides preliminary evidence that the use of LLMs can generate structurally appropriate antipersonas. The consistent performance observed for structural completeness and specificity suggests that combining role prompting with one-shot prompting is a viable strategy for automating the creation of this type of security artifact. The study findings also revealed recurring semantic inconsistencies, particularly the selection of formally valid but contextually inappropriate tactics and techniques. Overall, this work demonstrates that, when guided by carefully designed prompts and controlled technical documentation, LLMs can serve as effective assistants for antipersona generation, accelerating security analysis while complementing, rather than replacing, the expertise of security professionals.

It is important to emphasize that this study measures the internal quality of the generated artifacts and does not establish that the represented antipersonas correspond to real, plausible threats to OpenMRS; such validation is left for future work. Future work also includes evaluating the proposed approach across additional software systems and different LLMs.

Artifact Availability

The repository is available at https://anonymous.4open.science/r/ic-pibic-C088/README.md.

References

[1] Anthropic. 2026. Introducing Claude Sonnet 4.6. Anthropic News. https://www.anthropic.com/news/claude-sonnet-4-6

[2] Gheventer at al. 2026. Investigating the Use of Large Language Models for Generating Abuser Stories for Early Security Threat Identification.

[3] Andrea S. Atzeni, Cesare Cameroni, Shamal Faily, John Lyle, and Ivan Fléchais. 2011. Here's Johnny: A Methodology for Developing Attacker Personas. In *2011 Sixth International Conference on Availability, Reliability and Security (ARES)*. IEEE, 722–727. doi:10.1109/ARES.2011.115

[4] Martin Bedoya, Sara Palacios, Daniel Díaz-López, Estefania Laverde, and Pantaleone Nespoli. 2024. Enhancing DevSecOps practice with Large Language Models and Security Chaos Engineering. *International Journal of Information Security* 23 (2024), 3765–3788. doi:10.1007/s10207-024-00909-w

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.

[6] Badhan Chandra Das, M. Hadi Amini, and Yanzhao Wu. 2025. Security and Privacy Challenges of Large Language Models: A Survey. *Comput. Surveys* 57, 6 (2025), 1–39. doi:10.1145/3712001

[7] Isra Elsharef, Zhen Zeng, and Zhongshu Gu. 2024. Facilitating Threat Modeling by Leveraging Large Language Models. *Proceedings 2024 Workshop on AI Systems with Confidential Computing (AISCC)* (2024). doi:10.14722/aiscc.2024.23016

[8] Mavroeidis et al. 2021. Threat Actor Type Inference and Characterization within Cyber Threat Intelligence. In *2021 13th International Conference on Cyber Conflict (CyCon)*. IEEE, 327–352. doi:10.23919/CYCON51939.2021.9468305

[9] Schuller et al. 2024. Generating Personas Using LLMs and Assessing Their Viability. In *Extended Abstracts of the 2024 CHI Conf. on Human Factors in Computing Systems (CHI EA '24)*. ACM, Article 179, 7 pages. doi:10.1145/3613905.3650860

[10] Sahoo et al. 2024. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. arXiv:2402.07927

[11] Wimbauer et al. 2025. ThreatCompute: Leveraging LLMs for Automated Threat Modeling of Cloud-Native Applications. In *Proceedings of the 2025 Cloud Computing Security Workshop (CCSW '25)*. ACM, 14–27. doi:10.1145/3733812.3765533

[12] Shawn Hernan, Scott Lambert, Tomasz Ostwald, and Adam Shostack. 2006. Uncover Security Design Flaws Using The STRIDE Approach. *MSDN Magazine* (November 2006).

[13] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[14] Adam Steele and Xiaoli Jia. 2008. Adversary Centered Design: Threat Modeling Using Anti-Scenarios, Anti-Use Cases and Anti-Personas. In *Proc. of the International Conference on Information and Knowledge Engineering (IKE)*. 367–370.

[15] Blake E. Strom, Andy Applebaum, Doug P. Miller, Kathryn C. Nickels, Adam G. Pennington, and Cody B. Thomas. 2018. *MITRE ATT&CK: Design and Philosophy*. Technical Report. The MITRE Corporation.

[16] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*. doi:10.5555/3600270.3602070