

Article

Cyber Coercion Detection Using LLM-Assisted Multimodal Biometric System

Abdulaziz Almehmadi 

Department of IT, Faculty of Computing and IT, AIST Research Center, University of Tabuk,
Tabuk 47512, Saudi Arabia; aalmehmadi@ut.edu.sa

Featured Application

A cyber coercion detection system to identify user coerced actions.

Abstract

Cyber coercion, where legitimate users are forced to perform actions under duress, poses a serious insider threat to modern organizations, especially to critical infrastructure. Traditional security controls and monitoring tools struggle to distinguish coerced actions from normal user actions. In this paper, we propose a cyber coercion detection system that analyzes a user's activity using an integrated large language model (LLM) to evaluate contextual cues from user commands or actions and current policies and procedures. If the LLM indicates coercion, behavioral methods, such as keystroke dynamics and mouse usage patterns, and physiological signals such as heart rate are analyzed to detect stress or anomalies indicative of duress. Experimental results show that the LLM-assisted multimodal approach shows potential in detecting coercive activity with and without detected coercive communication, where multimodal biometrics assist the confidence of the LLM in cases in which it does not detect coercive communication. The proposed system may add a critical detection capability against coercion-based cyber-attacks, providing early warning signals that could inform defensive responses before damage occurs.

Keywords: cyber coercion; behavior analysis; physiological signals; biometrics; cybersecurity; large language models



Academic Editor: Stefan Fischer

Received: 27 August 2025

Revised: 30 September 2025

Accepted: 1 October 2025

Published: 2 October 2025

Citation: Almehmadi, A. Cyber Coercion Detection Using LLM-Assisted Multimodal Biometric System. *Appl. Sci.* **2025**, *15*, 10658. <https://doi.org/10.3390/app151910658>

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Rapid digital transformation across industries has significantly increased organizational dependence on information systems, consequently elevating cybersecurity threats and vulnerabilities. Cyber coercion, defined as manipulating or pressuring users into performing malicious or unauthorized actions, has emerged as a particularly insidious threat. According to Verizon's 2023 *Data Breach Investigations Report*, approximately 74% of cybersecurity breaches involve a human factor, often including social engineering, phishing, and coercive tactics [1]. These attacks pose severe risks, especially to critical infrastructures and sensitive information assets, and have seen an annual growth rate of 25% over the past three years as stated by a 2023 IBM security report [2]. While malicious insiders can sometimes be profiled through intent-driven behaviors, coercion-induced actions often mimic legitimate workflows. This makes coercion both more difficult to detect and potentially more damaging, as it bypasses traditional trust and authentication safeguards.

Traditional security measures have primarily relied on intrusion detection systems and reactive monitoring, yet these methods are insufficient for detecting subtle behavioral

cues indicative of coercion or undue psychological pressure. Recent studies highlight significant shortcomings in existing cybersecurity frameworks concerning internal threats involving coerced user behavior, as these rarely leave clear technical footprints and thus evade traditional detection mechanisms [3]. Cyber coercion can be considered a specific form of insider threat, where the insider does not act out of malicious intent but rather under external duress. Unlike malicious insiders who willingly abuse their access, coerced insiders are forced into compliance, making their actions appear legitimate to traditional monitoring systems. This overlap between insider threats and coercion underscores the urgent need for novel detection mechanisms capable of distinguishing voluntary from involuntary actions.

To address this gap, advanced analytics leveraging artificial intelligence (AI), specifically large language models (LLMs), offer promising solutions. Recent advances in AI, exemplified by OpenAI's GPT-4 and similar large language models like DeepSeek demonstrate sophisticated contextual understanding and anomaly detection capabilities [4]. However, to effectively detect coercion, these models require nuanced data about user behavior and physiological responses to accurately discern genuine threats from innocuous actions or decisions made under distress, especially when an LLM is aware of current company policies and procedures. In this context, LLMs play a unique role by interpreting user commands and communications for signs of coercion. For instance, an LLM can recognize when a command sequence appears externally instructed (e.g., unusual file deletions) or when written communication reflects stress or urgency inconsistent with normal workflows. This semantic awareness allows LLMs to detect cues that are invisible to rule-based anomaly detection systems.

The objective of this study is to design and evaluate a cyber-coercion detection system that integrates policy-aware large language model (LLM) analysis of user commands with behavioral (keystroke/mouse dynamics) and physiological (heart rate/HRV) signals in order to detect situations where users may be acting under duress.

This research proposes an innovative approach integrating behavioral biometrics, such as keystroke dynamics and mouse movement, and physiological biometrics, including heart rate monitoring through camera-based photoplethysmography, with large language models. By merging these technologies, the proposed system aims to identify coercion by detecting abnormal patterns indicating psychological stress or external pressure.

Given the high stakes associated with cyber coercion, especially in sensitive sectors such as finance, healthcare, and national security, the development and deployment of such proactive detection systems are critically important. To illustrate this challenge in practical terms, consider a system administrator who receives a threatening message instructing them to delete critical organizational databases. Under duress, the administrator proceeds with the deletion. From a system log perspective, these commands may appear to be routine administrative tasks.

However, an LLM analyzing the textual context of the threat message and the atypical sequence of commands could flag this as potential coercion. When combined with behavioral indicators such as faster typing with errors and physiological signs of stress (elevated heart rate), the system achieves a higher confidence in detecting coercion. This scenario demonstrates how contextual reasoning by LLMs, supported by multimodal biometrics, provides an effective defense against coercion-driven insider threats. Behavioral and physiological biometrics were chosen as the focus of this research because they provide strong, complementary indicators of user stress and duress, which are hallmarks of coercion.

Prior studies have demonstrated that stress can be reliably detected through keystroke dynamics and mouse movement anomalies, as well as physiological responses such as elevated heart rate variability (HRV). These modalities are particularly valuable because

they can be captured continuously and non-intrusively during normal system use, providing a practical layer of evidence to support LLM-based contextual reasoning. Together, they enable a multimodal approach that increases robustness and confidence in detecting coercion-driven insider threats.

To the best of our knowledge, this is the first work to integrate policy-aware LLM reasoning with multimodal behavioral and physiological biometrics specifically for detecting coercion-based insider threats. While prior studies have explored biometrics or LLMs separately, our framework uniquely fuses both modalities to strengthen detection confidence; therefore, this paper investigates the feasibility and effectiveness of combining behavioral and physiological biometric analyses with advanced LLMs for robust, cyber coercion detection.

The scope of this paper focuses on proposing a coercion detection system that integrates three critical components: behavioral and physiological biometrics, and large language model (LLM)-based contextual reasoning. The focus is on detection rather than on designing full defensive countermeasures, although accurate detection is a necessary prerequisite for effective organizational defense.

The remaining of this paper is organized as follows: In Section 2, we present related research. In Section 3, we present the proposed system design and methodology. In Section 4, we present the results, and we finally provide conclusions and plans for future work in Section 5.

2. Related Work

Cybersecurity threats are continuously evolving, becoming more sophisticated, targeted, and often deeply personalized. Recent advancements in cybersecurity reveal a significant rise in insider threats and coerced actions executed by legitimate users who have authorized access. Insider threats are among the hardest to detect since they originate within an organization's boundaries and utilize legitimate user credentials. According to the 2023 IBM Security X-Force report [2], insider threats accounted for 29% of all cyber incidents, making them one of the most critical areas requiring immediate and effective detection methods. Insider threats based on subject coercion have been studied from various angles including behavioral analysis of subjects and physiological signals.

2.1. The Role of Behavioral and Physiological Biometrics in Stress Detection as an Indication of Cyber Coercion

Security researchers and human-computer interaction experts have long investigated how stress and other cognitive states can be identified through user behavior. Behavioral biometrics such as keystroke dynamics and mouse movement patterns provide continuous and unobtrusive monitoring of a user's interactions. For example, the authors of [5] explored detecting cognitive and physical stress by analyzing free-text typing and found it feasible to classify stress versus non-stress conditions using keystroke timing and linguistic features, achieving accuracy comparable to dedicated physiological sensors. The attractiveness of such an approach is that it requires no additional hardware beyond a standard keyboard, making it low-cost and transparent to the user. Subsequent studies reinforced these findings. The authors of [6] provided a survey reviewing sensor-based, unobtrusive methods including physiological, behavioral, and environmental data for continuous stress monitoring in knowledge-work environments, highlighting real-world applications, user acceptance factors, and the challenges in achieving personalized, long-term adoption. The authors of [6] used a pressure-sensitive keyboard and a capacitive mouse to sense stress in a controlled user study. They reported that, under stressful conditions, over 79% of participants exhibited significantly increased typing pressure and 75% showed greater

contact with the mouse surface. These behavioral changes align with the physiological “fight-or-flight” response such as elevated heart rate and muscle tension manifesting in computer interactions. Such results demonstrate that typing behavior and pointer movements can serve as proxies for physiological stress. In another study by the authors of [7], the authors focused on smartphone typing behaviors and motion sensors, achieving up to 87.5% accuracy in distinguishing stress vs. calm states using accelerometer and gyroscope data from typing sessions.

Further, behavioral analysis has emerged as a promising approach to identifying and mitigating insider threats. Asasfeh et al. [8] presented a comprehensive study on insider threats, categorizing detection techniques into signature-based, anomaly-based, and hybrid approaches. Their study emphasized the importance of monitoring deviations from established behavioral baselines, such as typing speed, access patterns, and resource utilization, to identify suspicious activities. Behavioral biometrics, including keystroke dynamics and mouse movement, have gained attention due to their effectiveness in differentiating between authorized users and potential threats [9–12].

Furthermore, in recent years, physiological biometrics have been introduced into cybersecurity frameworks to enhance user authentication and anomaly detection. Physiological signals such as heart rate variability, facial expressions, and eye-tracking have shown promising results in detecting stress and emotional changes that may indicate coercion or distress [13,14]. For instance, research conducted by the authors of [15] demonstrated that heart rate variability can reliably indicate stress levels in cybersecurity tasks, highlighting its potential as an effective biometric marker for coercion detection.

This illustrates the efficacy of cross-modal behavioral features for stress detection across different devices. Researchers have also investigated physiological biometrics for implicit stress sensing, such as heart rate, skin conductance, or pupil dilation. These signals can be highly indicative of stress, but often require dedicated hardware (wearables, eye trackers, etc.) and may introduce user discomfort. Therefore, recent work tends to favor behavioral indicators or minimal sensors that can be seamlessly integrated. Non-invasive stress detection via keystroke dynamics is a particularly active area because keyboard data is readily available in many cybersecurity contexts such as employees typing commands or messages. The literature collectively indicates that by monitoring typing speed, latency variations, error rates, and pressure surrogates, one can infer a user’s stress level or emotional state with reasonable accuracy. These insights lay the foundation for using behavioral and physiological cues as part of a coercion detection system, since a user under coercion is likely to be under stress or exhibit atypical interaction patterns.

2.2. Coercion Detection

Detecting coercion has been acknowledged as a crucial but difficult problem in biometric security for at least a decade. The authors of [16] were among the first to formally outline the need for coercion detection mechanisms in biometric systems. They identified scenarios such as an attacker forcing a victim to authenticate, where traditional biometrics fail to distinguish voluntary from involuntary login attempts. Following this, research efforts began exploring technical solutions to this challenge such as the study conducted by [17].

The authors of [18] proposed an innovative approach to improve biometric authentication by integrating liveness detection to ensure the user is physically present and coercion detection to ensure that the user is not an artificial spoof. The objective was to develop an algorithm that assesses multiple aspects of a biometric attempt to determine if the user is not only live but also acting willingly. By fusing liveness and coercion checks, their method aimed to enhance the overall robustness of authentication systems against both spoofing

and duress attacks. One strategy in coercion detection for authentication is to define duress indicators or codes. In some security systems such as bank vaults or ATMs, a user can enter a special duress PIN that still unlocks the system but also silently signals distress to authorities.

However, expecting users to remember and use duress signals is error-prone and not applicable in many cyber scenarios. Therefore, research has gravitated towards implicit detection via behavioral cues. For instance, if a user is being coerced during login, they might have abnormal hesitation or erratic input patterns despite eventually entering correct credentials. Incorporating behavioral biometric analysis as suggested by the above works can provide these implicit signals.

Our research paper [18] proposed fingerprint placement analysis in milliseconds that showed high potential in differentiating between willingness to provide a fingerprint when compared to being forced to place a fingerprint. Another angle is anomaly detection on user activity patterns. If an employee suddenly performs atypical tasks at someone else's direction, such as accessing files they never usually handle under the coercer's instruction, this could be detected by user behavior analytics. However, pure anomaly detection might not discern coercion from other causes of anomaly, such as malicious insider intent. What sets dedicated coercion detection apart is the focus on recognizing signs of duress in the user's behavior at the time of the action. Prior studies like [19] leveraged biometric readings available alongside the primary authentication biometric to decide if the user is under threat.

2.3. Large Language Models in Cybersecurity

The past few years have seen rapid advances in large language models and their application beyond traditional NLP tasks. In cybersecurity, researchers are beginning to utilize LLMs to detect anomalies and malicious intents by analyzing unstructured data such as text logs, commands, and emails in a more context-aware manner. Unlike rule-based systems, LLMs can interpret subtle cues and learn complex patterns from data. The authors of [20] demonstrate this capability in the domain of network intrusion detection: they applied a BERT-based model to network event streams and achieved near-perfect accuracy in identifying attacks in IoT network traffic, significantly outperforming classical detection methods. This suggests that pre-trained transformer models can capture the nuances of "normal" vs. "malicious" behavior in sequences of events, a principle that could equally apply to user activity sequences. Moreover, LLMs have shown proficiency in analyzing the content of communications for signs of deceit or mal-intent.

A recent extensive survey by [21] on LLM security applications reports that prompted LLMs were able to detect malicious intent in phishing emails with high effectiveness, in some cases even surpassing human detection performance. The survey notes experiments where models like GPT-4 and others correctly flagged phishing or malicious requests that humans missed, indicating the models' potential in recognizing subtle indicators of coercion or fraud in language. If an employee under coercion is communicating with an attacker such as via chat or email instructions or even talking to themselves such as writing unusual command comments, an LLM could analyze those text patterns for stress or coercive language.

LLMs can also dynamically understand context, for example, distinguishing a normal administrative command from one that is out-of-character for the user's role, when combined with semantic knowledge of the organization's operations. These developments in AI-driven security monitoring inform our approach. By incorporating an LLM into the coercion detection framework we propose, we aim to leverage its strength in contextual and semantic analysis with specific policies and procedures. Traditional anomaly detection

might raise an alert on unusual file access, but an LLM-based analysis might further interpret why it is unusual, such as the content of a command sequence suggesting the user was following someone else's instructions.

As shown by the above studies, LLMs can enhance anomaly detection across domains and applying them to interpret user behavior and language in security logging is a logical next step. Our work is, to our knowledge, the first to integrate LLM analysis with multimode based biometric stress indicators specifically for detecting coerced actions in a cyber environment.

2.4. Gaps in the Current Research

Despite significant advancements, the existing literature exhibits clear gaps, particularly regarding practical systems capable of detecting coercion based on combined behavioral and physiological biometrics analyzed by AI-driven language models. Most studies tend to focus separately on either behavioral biometrics, physiological biometrics, or NLP-driven analyses. Few studies provide integrated models capable of holistically interpreting multiple data streams to detect coercion, especially under high-stress conditions.

The current literature also lacks comprehensive frameworks that explicitly consider the trade-off between the immediate consequences of actions and their long-term impacts, particularly in scenarios involving coercion. For instance, current anomaly detection systems typically respond to immediate threats without contextualizing the potential consequences of user actions or the motivations behind them, potentially misclassifying beneficial user actions conducted under stressful conditions as malicious. This gap underscores the need for an integrated framework as proposed in this research, leveraging LLMs to interpret complex scenarios, physiological biometrics to gauge emotional and psychological states, and behavioral biometrics to track action consistency. Such an integrated system would provide nuanced insights into user intentions and motivations, significantly enhancing the capability to detect genuine coercion scenarios versus legitimate yet risky actions. Table 1 summarizes key related work, the methods used, the advantages and the limitations.

Table 1. Summary of key related work, the method used, the advantages, and the limitations.

Reference	Method Used	Advantages	Limitations
Vizer et al. [5] (2009)	Keystroke and linguistic features for stress detection	Non-intrusive, no special hardware needed, cost-effective	Limited accuracy, linguistic context dependency
Hernandez et al. [6] (2014)	Pressure-sensitive keyboard and capacitive mouse	High detection accuracy (79–75% sensitivity)	Specialized hardware required, not commonly available
Sagbaş et al. [7] (2020)	Smartphone sensors (accelerometer and gyroscope)	High accuracy (~87.5%) using widely available mobile sensors	Limited applicability to desktop environments, requires mobile device use
Asasfeh et al. [8] (2024)	Behavioral anomaly detection for insider threats	Comprehensive categorization, effective baseline deviation detection	No physiological integration, higher false positives
Ali et al. [9] (2008)	Automated multi-agent behavioral monitoring system	Continuous monitoring capability, early threat detection	Potential user privacy concerns, increased complexity
Almomani et al. [10] (2024)	Facial expression and behavioral biometrics	Effective proactive detection of insider threats	Requires facial recognition technology, privacy issues

Table 1. Cont.

Reference	Method Used	Advantages	Limitations
Arrabito et al. [13] (2024)	Heart rate variability (HRV) for stress detection	Reliable physiological biomarker for stress indication	Requires dedicated physiological sensors, intrusive setup
Matthew & Anderson [16] (2016)	Biometric coercion detection methodologies	Early identification of coercive authentication	Difficulty distinguishing voluntary vs. involuntary biometric entries
Almehmadi [18] (2022)	Fingerprint placement time analysis	Rapid, accurate differentiation of coerced authentication	Limited to fingerprint biometric context
Matthew & Canning [19] (2020)	Liveness and coercion detection fusion algorithm	Robust dual protection against spoofing and coercion attacks	Complexity in fusion logic, potentially high false alarms
Maasaoui et al. [20] (2024)	Large language models (LLM) using BERT-based transformer on network event streams (IoT traffic)	High semantic accuracy, near-perfect detection in IoT environments	Dependent on data quality, complex model tuning required
Yao et al. [21] (2024)	LLM-based phishing email and malicious intent detection	Superior detection performance over human analysts	Potential bias from training data, high computational costs

3. System Design and Methodology

3.1. System Architecture Overview

The proposed system is composed of three main modules working in tandem: (1) Data Collection and Preprocessing, (2) Coercion Detection, and (3) Decision and Alerting.

The Data Collection module continuously monitors the user's interactions, commands written, files opened/sent, and physiological signals such as heart rate. It records current activities, keystroke events with timestamps for each key press and release to derive timing features, mouse movements and clicks, and heart rate from a camera for non-intrusiveness. The raw data is preprocessed into features. Current undertaken activities are recorded and keystroke timing is converted into metrics like inter-key intervals, typing speed, error rate, backspaces per character, and latency patterns. Mouse data yields features such as movement speed, idle time between actions, click frequency, and angle of mouse movement. These form the behavioral feature vector. In parallel, the system computes a physiological stress feature from heart rate using a camera. The camera-based photoplethysmography (PPG) technique has been shown in prior work to provide reasonably accurate heart rate estimation under stable conditions, though accuracy may degrade with subject movement or poor lighting. In cases where a camera is unavailable or where organizational policy restricts its use, the system continues to function using behavioral biometrics (keystroke dynamics and mouse interactions) and LLM-based contextual analysis. This allows the system to maintain detection capability, albeit with potentially reduced confidence, thereby ensuring operational continuity without relying solely on physiological inputs [22]. All features are timestamped and synced along with the current activity the user is undertaking by recording the network traffic and system logs. User interactions are compared against a database composed of current company policies and procedures using the LLM to determine if a user is acting against the rules. If performing the activity is not against the rules, the LLM optimizes itself to reduce false positives and false negatives over time by continually learning. If, on the other hand, an action that is against the rules is detected, the Coercion Detection module is activated.

The Coercion Detection module consists of a multimodal biometric, LLM-based module and one-class SVM. First, the full data, including behavioral and physiological signals prior and during the detected wrongful act, is fed to the one-class SVM, which was trained on normal user data to flag statistically significant anomalies in user keystroke, mouse patterns, and heart rate. Next, an LLM-based context analyzer processes textual context in parallel including logs, network traffic, and email communication to form the context, taking into consideration the policies and procedures of the company and the result from the one-class SVM and the biometric data. This context can include the actual commands the user is executing or the content of chats/emails if the coercion involves communication. We fine-tuned a distilled DeepSeek model (DeepSeek-R1-Distill-Qwen-32B) [23], on a small artificially made corpus of legitimate vs. coerced task descriptions as stated in Section 3.2.

The LLM analyzes each new command or message, producing a score indicating how unusual or coerced it sounds given the user's typical profile and general semantic indicators of duress such as unusually terse or error-laden command sequences and language that suggests urgency or external instructions. Then, manual validation is used to optimize the LLM and the cycle repeats.

The Decision and Alerting module will, upon a positive detection, log the event and raise an alert to security operators. In a real deployment, this module could also automatically pause the user's session or require re-authentication using a secret panic code. In our experimental setup, we focused on logging the detections for offline analysis. The entire system is designed to operate with minimal latency; feature extraction and model inference run continuously with negligible delay. The LLM analysis is the heaviest component, but by using a distilled model and limiting input length, it processes each command in under 1 s given the limited resources on our hardware. This ensures that if a coercion is detected, an alert can be raised almost immediately once the hardware limitation is addressed, potentially fast enough to interrupt the coerced transaction. Figure 1 depicts the proposed architecture for the LLM-assisted multimode biometric system for detecting coerced actions. Collaboration among the three modules ensures a layered and complementary defense mechanism.

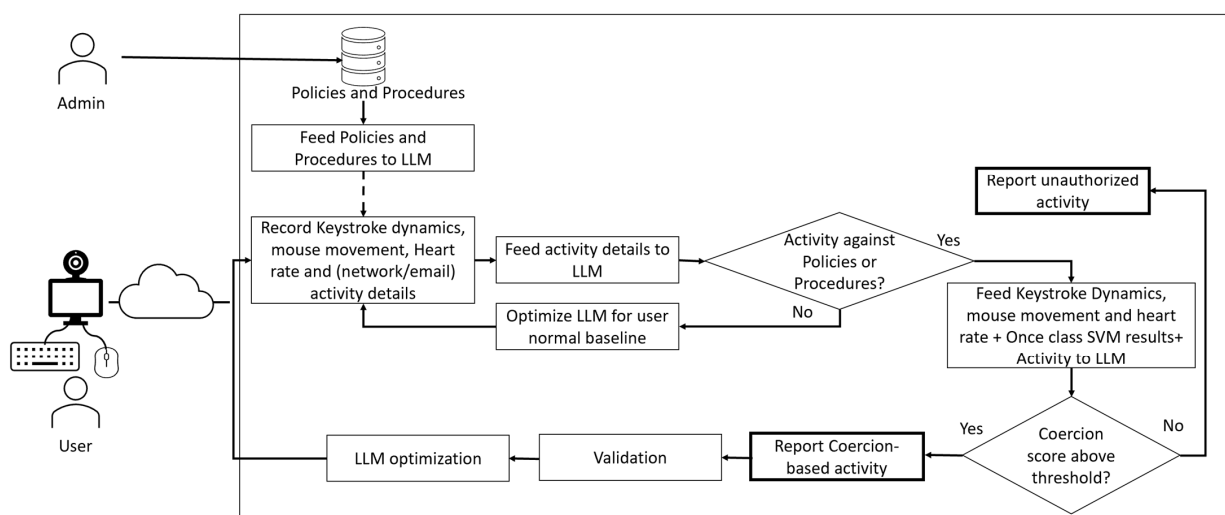


Figure 1. The proposed system architecture integrating LLM and multimodal biometrics to detect coercion.

The Data Collection and Preprocessing module continuously streams synchronized behavioral, physiological, and contextual system activity data, which form the foundation for subsequent analysis. These features are then passed to the Coercion Detection module, where

the one-class SVM statistically identifies anomalies, and the LLM performs semantic and policy-aware reasoning. This integration allows the system to combine low-level statistical deviations with high-level contextual interpretation, thereby minimizing false positives and enhancing detection accuracy. Finally, the Decision and Alerting module aggregates these insights and translates them into actionable outcomes, such as generating alerts, logging incidents, or requiring re-authentication. Each module complements the others: data collection ensures completeness, coercion detection ensures interpretability and robustness, and decision-making ensures operational responsiveness. Collectively, these architectures create a synergistic framework that improves reliability and timeliness in network threat detection, enabling rapid intervention in coercion-based cyber-attacks. Figure 2 depicts the detailed biometric data before they are fed to the LLM with the one-class SVM data and activity.

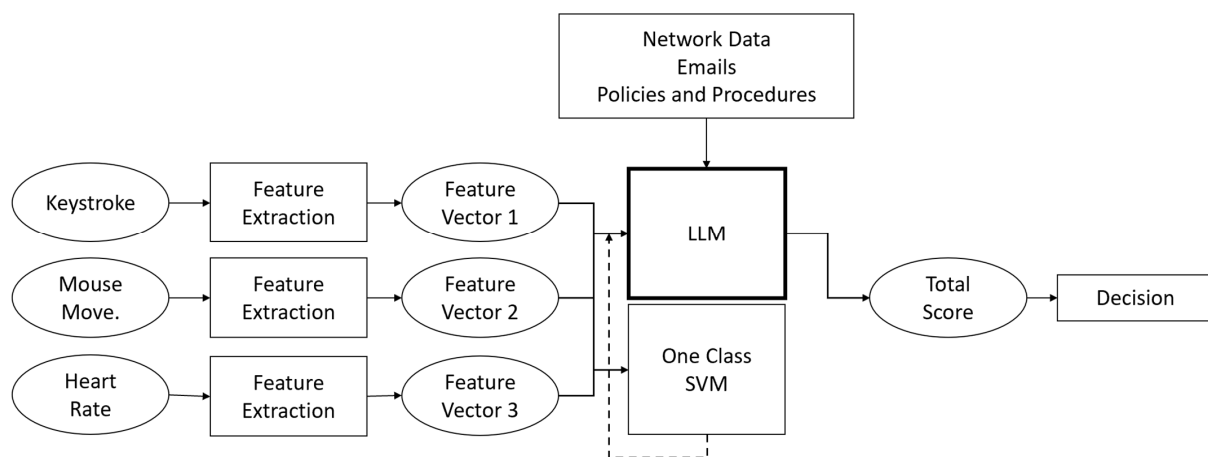


Figure 2. The proposed system detailing the biometric data and one-class SVM results to be fed into the LLM to detect coercion based on LLM total score.

Full implementation steps, feature definitions, model hyperparameters, and prompts required to reproduce the experiments are detailed in Section 3.2.2 (“Reproducibility and Open Materials”) and the Supplementary Materials.

3.2. Implementation Details

The proposed system was implemented in Python 3.10, using PyTorch 2.0 for biometric modeling and the DeepSeek-R1-Distill-Qwen-32B for contextual analysis. Preprocessing steps included normalization, segmentation into 60 s windows, and extraction of keystroke/mouse features such as inter-keystroke interval variance and mouse trajectory deviation. For physiological signals, heart rate (HR) and heart rate variability (HRV) were derived from camera-based photoplethysmography. The LLM component was fine-tuned on a custom dataset. Prompts were designed to simulate company policy contexts and typical coercion scenarios. The LLM output was scored on whether the activity aligned with policy-consistent or coercion-indicative behavior. Integration between the biometric modules and the LLM was achieved through a decision fusion layer, where biometric anomaly scores and LLM contextual scores were combined using weighted averaging. All experiments were conducted on a workstation with an Intel i9 processor, 32 GB RAM, and NVIDIA RTX 3090 GPU. A step-by-step execution view is given in the algorithmic workflow pseudocode below, which details training, inference, fusion, and alerting.

Fusion Weight and Threshold Calibration

To optimally combine the OC-SVM anomaly score (A) and the LLM coercion score (L), we employed a weighted fusion scheme defined as $F = wA \cdot A + wL \cdot (L/100)$. Candidate weights (wA , wL) were selected through grid search on the validation set to maximize the

area under the ROC curve (ROC-AUC). The final chosen configuration slightly favored the LLM score ($w_L > w_A$) to reflect the higher reliability of contextual evidence when available. The alert threshold (T_{alert}) was determined using Youden's J statistics, with additional verification through F1 optimization to ensure balanced sensitivity and specificity. This process yielded a stable threshold across validation folds, reducing false positives in non-coercive but stressful workplace scenarios.

Statistical Significance Testing

To quantify robustness, we performed non-parametric bootstrap resampling (1000 iterations) over task-level outcomes. This produced 95% confidence intervals for accuracy, ROC-AUC, and other reported metrics (The results confirm that multimodal fusion significantly outperformed unimodal baselines ($p < 0.05$, bootstrap test), underscoring the value of integrating behavioral, physiological, and contextual features.

Because stress can arise from routine work (deadlines, multitasking, etc.), we separate stress detection from coercion attribution. First, we detect stress/anomaly using the biometric model (keystroke, mouse, and HR/HRV). Second, we require contextual corroboration, identified by the policy-aware LLM, such as (1) explicit coercive language in messages, (2) policy/permission violations or out-of-role actions in commands/logs, or (3) atypical command sequences inconsistent with the user's role. A "coercion" label is issued only when stress/anomaly co-occurs (within the 60 s window, ± 30 s) with at least one contextual indicator; otherwise, the event is labeled non-coercive workplace stress.

Algorithmic Workflow (Pseudocode) (Algorithms 1 and 2):

Algorithm 1. Training And Calibration

```
# Feature extraction and normalization
  Build synchronized, 60 s windows with 30 s overlap from D_norm (drop
  degenerate windows).
  Extract window-level features:
    Keystroke: dwell, flight (IKI) stats, error/backspace ratios, bursts, speed.
    Mouse: speed/accel stats, idle-time ratio, click freq, angle-change rate,
    tortuosity, hesitations.
    PPG/HR: HR mean, SDNN, RMSSD, pNN50, LF/HF (fallback to time-domain
    if quality low).
  Per-user z-score normalize features.

# Train OC-SVM on normal behavior
  Grid-search  $\nu \in \{0.01, 0.05, 0.1\}$ ,  $\gamma \in \{1/d, 0.1/d, 10/d\}$  with 5-fold CV on normal
  windows only.
  Fit RBF OC-SVM  $\rightarrow$  obtain anomaly score  $A \in \mathbb{R}$ ; min-max scale A to  $[0, 1]$  on
  validation.

# Fine-tune LLM on context/policy data
  Construct instruction-style pairs: (policy snippets + action/comm. traces)  $\rightarrow$ 
  {policy-consistent, coercion-indicative}.
  Fine-tune base model (DeepSeek-R1-Distill-Qwen-32B) with fixed hyperparameters.
  Define inference template returning {coercion: bool, score: 0–100}.

# Calibrate fusion and threshold
  For validation windows: compute OC-SVM score A and LLM score L.
  Fuse  $F = w_A \cdot A + w_L \cdot (L/100)$ , choose  $w_A, w_L$  by maximizing ROC-AUC.
  Pick  $T_{\text{alert}}$  using Youden's J (or F1) on validation ROC/PR curves.
  Persist  $\{\theta_{\text{svm}}, \phi, w_A, w_L, T_{\text{alert}}\}$ .
```

Algorithm 2. Online Detection

Collect current 60 s window of keystroke, mouse, (optional) PPG/HR; synchronize with commands/logs/emails.

Extract and normalize features as in Algorithm 1 (fallback gracefully if a modality is missing).

$A \leftarrow$ OC-SVM anomaly score in $[0, 1]$ using θ_{svm} .

$L \leftarrow$ LLM coercion score $\in [0, 100]$ from φ using the inference prompt with policy/context.

$F \leftarrow w_A \cdot A + w_L \cdot (L/100)$.

If $F \geq T_{\text{alert}}$:

 Trigger alert: record context, scores, and policy references; (optionally) require re-authentication.

Else:

 Continue monitoring.

Apply hysteresis: require k consecutive positive windows to reduce flicker (k configurable).

3.2.1. Camera-Based PPG Robustness and Alternatives

Camera-based photoplethysmography (rPPG) is sensitive to motion and illumination changes. To mitigate this, we apply lightweight quality checks per window (face/ROI-tracking stability, global illumination variance, and rPPG SNR heuristics). Windows failing these checks are marked PPG-unreliable: we then (i) fall back to time-domain HR/HRV only when spectral features are unstable or (ii) drop physiological features entirely and proceed with behavior + LLM context (reduced physiology weight in fusion). The pipeline is modality-agnostic: when wearable signals (e.g., wrist PPG or chest ECG/EDA) are available, they can replace or augment rPPG after re-calibrating fusion weights and thresholds.

3.2.2. Reproducibility and Open Materials

Experiments were run in Python 3.10 with PyTorch and scikit-learn on a workstation (RTX 3090 GPU, Intel i9 CPU, 32 GB RAM). Keystroke, mouse, and PPG/HR streams were synchronized to a common timeline and segmented into 60 s windows with 30 s overlap. Windows with fewer than ten keystrokes or fewer than two mouse trajectories were discarded. Features were then standardized per user using a rolling baseline of normal activity to emphasize within-subject deviations.

Feature families comprised keystroke timing statistics, mouse movement/click summaries, and HR/HRV measures from camera-based PPG (with a fallback to time-domain HR/HRV when spectral quality was poor). A one-class SVM was trained on normal windows to produce a normalized anomaly score. The LLM component (DeepSeek-R1-Distill-Qwen-32B) was fine-tuned to provide a policy-aware coercion score from commands/logs/emails plus biometric summaries. The decision layer combines the anomaly and LLM scores with a slightly higher weight on the LLM to reflect contextual evidence. The alert threshold was selected on validation data to balance sensitivity and specificity. We report precision, recall, F1-score, and ROC-AUC in Section 4.1. To quantify variability across tasks/conditions, we use non-parametric bootstrap resampling (1000 resamples) over task-level results per scenario. Bars and point estimates in figures report the mean with 95% confidence intervals (CIs). We also compute CIs for aggregate metrics such as Accuracy/ROC-AUC by resampling task outcomes; the figure error bars correspond to these 95% CIs.

To keep the main text focused, full code-level steps, feature definitions, hyperparameter grids, and prompt templates are provided in the Supplementary Materials.

3.3. Data Collection and Simulation-Based Experimental Setup

Due to significant ethical concerns associated with conducting coercion experiments on individuals, the study adopted a simulation-based approach. This approach allows an indication of the effectiveness of the proposed system and a start evaluation of the coercion detection system while strictly adhering to ethical guidelines.

3.3.1. Normal User Behavior Data Collection

To establish a baseline for normal user behavior, we constructed a simulated dataset that reflects typical administrative tasks without coercion. A single interaction sequence was first recorded, consisting of three categories of tasks: (1) updating user account information, (2) deleting user accounts, and (3) deleting a database. This sequence served as the foundational template for normal system usage. From this template, 20 distinct baseline scenarios were generated to emulate repeated executions of the same tasks by different individuals. To ensure that the dataset captured natural variability in human behavior, controlled perturbations were introduced into each scenario. Specifically: Keystroke dynamics were varied by applying small Gaussian noise to keypress and release intervals, simulating differences in typing speed and rhythm. Mouse movement patterns were modified by adjusting cursor trajectories, introducing micro-pauses, and varying click delays and frequencies. Error behaviors such as backspaces and typing corrections were inserted probabilistically at low rates (1–3%) to mimic realistic human typing variability. These controlled modifications ensured that each scenario maintained the same overall task sequence but exhibited slightly different interaction patterns, reflecting inter-user differences. The resulting dataset provided a balanced baseline of “normal” activity against which coercion-indicative signals could be integrated. To simulate coercion, behavioral and physiological stress responses were later incorporated from validated public datasets (SWELL-KW and WESAD). These datasets include diverse participants across gender and age, thereby embedding natural variation in stress responses. The integration of these external signals allowed the system to be evaluated across multiple coercion scenarios while avoiding ethical risks associated with conducting genuine coercion experiments.

Baseline Scenario Synthesis Parameters

Starting from one recorded interaction sequence spanning the three task categories (update user, delete user, and delete database), we created 20 baseline scenarios by applying controlled, bounded perturbations while preserving task order. For keystrokes, inter-keystroke intervals were jittered with zero-mean Gaussian noise whose spread was set to about 8% of the per-user median interval, then clipped to 30–800 ms. Dwell times were jittered with zero-mean Gaussian noise at ~6% of the per-user median dwell and clipped to 40–350 ms. To mimic natural errors, we inserted backspaces/typing corrections in 1–3% of characters.

For mouse behavior, we added small frame-wise spatial jitter (about 1.5 pixels at 60 Hz). Instantaneous cursor speed was randomly scaled with a modest spread (about 15%), then clipped to the 1st–99th percentile of the original distribution. We injected micro-pauses roughly once every 15 s, with durations centered around 200 ms (minimum 120 ms), and added small random delays to clicks (standard deviation ~20 ms, clipped to non-negative values). These choices introduce realistic variability without distorting overall path geometry.

All perturbations were followed by per-user z-normalization to emphasize within-subject deviations while preserving timing structure. Each synthesized scenario used a fixed random seed to enable exact regeneration.

3.3.2. Integration of Coercion-Indicative Signals from Validated Datasets

To simulate realistic coercion scenarios without direct ethical risks, behavioral and physiological signals indicative of stress and coercion were integrated from established and validated research datasets. Specifically, the following datasets were utilized: the SWELL Knowledge Work (SWELL-KW) dataset for stress-induced keyboard and mouse behavior [24] and the WESAD dataset containing ECG and PPG physiological signals recorded under controlled stress scenarios [25]. The stressed behavioral and physiological signals were assigned to all 3 categories of tasks to evaluate how the system will react. We understand that behavioral and physiological signals may be thought to impact the merits of the data analysis and system evaluation; however, the scope of the paper is to measure the effectiveness of the LLM-assisted multimodal biometric system, the proposed system, in its ability to detect coercion signals, integrated stress signals, and the contexts of different tasks to provide an alert; in particular, we aim to test the system with and without stress signals and report the findings.

Privacy and Ethical Considerations

Given the sensitivity of behavioral and physiological monitoring, particular attention was paid to privacy and ethics. First, all camera-based PPG signals were processed locally on-device, and no raw video streams were stored or transmitted. Only anonymized physiological features (e.g., HR and HRV measures) were retained for analysis. Second, no real coercion experiments involving human participants were conducted in this study. Instead, stress-related data were integrated from publicly available, ethically reviewed datasets (SWELL-KW and WESAD), both of which were collected under institutional approval and with informed consent from participants. Consequently, our simulation-based evaluation avoided exposing new participants to coercion scenarios. Finally, we acknowledge that future real-world deployments will require rigorous institutional ethics review, strong privacy-preserving mechanisms (such as on-device feature extraction, secure data handling policies, and minimal data retention), and organizational transparency to ensure responsible use of coercion detection technologies.

3.3.3. Procedure for Integrating Coercion Signals

The integration of coercion signals into normal interaction data followed a structured and replicable procedure designed to preserve temporal realism and consistency, which is explained as follows:

1. Behavioral Signal Integration (Keystroke and Mouse Dynamics):

Segments of keystroke and mouse interaction data indicating stress such as increased typing speed, irregular keypress intervals, jittery mouse movements, and accelerated cursor speeds were extracted from the SWELL-KW dataset. These stress-induced segments replaced original, neutral segments in the collected normal data at carefully selected interaction points.

2. Physiological Signal Integration (Heart Rate):

Corresponding physiological stress data segments from the WESAD dataset were aligned temporally with the integrated behavioral stress segments. PPG data reflecting increased heart rate and heart rate variability (HRV) changes which are known markers of acute stress and coercion were synchronized precisely to the behavioral integrations. Specifically, elevated physiological responses were inserted starting slightly before the simulated coerced behavioral events and continued through the coerced interaction period, gradually returning to baseline post-event, realistically mimicking human physiological response to coercive stressors.

3. Temporal and Contextual Alignment:

All integrated data segments were matched in duration and contextual relevance such as complexity and urgency to their corresponding original data segments to maintain realistic interaction flow. This alignment ensured that coercion signals matched realistic user–task interactions without causing unnatural interruptions or anomalies that might bias the detection outcomes.

4. Data Labeling:

Clear labeling was applied at each integration interval indicating coercion vs. normal activity periods. These labels will enable the evaluation of the detection system's accuracy and allow detailed analysis of both false positive and false negative detection events. Through this integration process, the final dataset provided a semi-realistic simulation environment, effectively suitable for evaluating the coercion detection model's performance especially at this initial stage of testing the system.

5. Data Integration Justification

It is recognized that integrating real but simulated coercion indicators into collected datasets constitutes intentional data manipulation. However, this intentional integration is justified within the scope of our research, given the ethical impossibility of inducing genuine coercive situations. The primary objective of this intentional data integration is to rigorously evaluate the capability and accuracy of the integrated large language model (LLM)-assisted multimodal biometric coercion detection system. By embedding carefully selected coercion signals from validated datasets, we explicitly create known coercive scenarios. This process allows for precise evaluation of the LLM's performance, specifically its capacity to: (1) Identify the presence and context of coercion by determining whether coercive behavior is occurring based on integrated behavioral, physiological, and contextual cues. (2) Evaluate the severity and urgency levels by assigning scoring based on the nature and magnitude of observed anomalies aligned with organizational policies and procedures. (3) Correlate multimodal biometric indicators by cross-referencing findings from behavioral analyses, physiological stress indicators, and anomaly detection outcomes from models such as the one-class SVM as well as email markers that signal coercive activity. Thus, the deliberate integration of coercion signals is essential for systematic assessment, ensuring the reliability, robustness, and contextual sensitivity of the coercion detection system under known and controlled conditions. It serves as the first step in evaluating the capabilities of the system, but it is by no means a final and full real-life evaluation system.

6. LLM Prompt Design

A carefully structured prompt was provided to the LLM to ensure precise contextual interpretation of system activities. The prompt explicitly defined parameters for recognizing coercion based on policies, procedures, and multimodal biometric analyses. The structured prompt is as follows: "Act as an expert security analyst. Given the recorded user interaction data including keystroke dynamics, mouse behaviors, physiological signals and organizational context derived from integrated security policies and operational procedures, and the systems logs and network traffic and emails if any, evaluate the following user actions. Identify and assess the likelihood and severity of coercion. Provide a detailed context-based analysis, referencing relevant policies and procedures, and assign an appropriate coercion risk score. Utilize the one-class SVM results to make the decision if data is available. The expected output is if coercion exists or not and if yes, how confident you are, provide a score from 0–100". This explicit prompt facilitates consistent and detailed analysis by the LLM, ensuring thorough evaluations grounded in both biometric data and

organizational guidelines. The result score can then be used to raise an alert if coercion is detected based on a predefined threshold.

7. Policies and procedures

To ensure contextually accurate evaluations by the LLM and reflect realistic organizational environments, the following carefully selected policies and procedures were integrated into the LLM derived from ISO/IEC 27001 and NIST:

a. User Account Management Policy:

It clearly defines permissible actions, authority levels, and required authorization for critical account activities such as creating, updating, or deleting user accounts. Violations indicate potential coercion or malicious actions.

b. Data Access and Authorization Policy:

It specifies appropriate access controls and user permissions. Unauthorized deviation or misuse of granted privileges signifies possible coercive or malicious intent.

c. Incident Response Policy:

It describes procedures for handling cybersecurity incidents, including coercion indicators or abnormal user behaviors, specifying immediate steps for reporting, documenting, and responding to coercion-related alerts.

3.4. Dataset Designed Scenarios for System Evaluation

To comprehensively evaluate the system's capabilities, we designed multiple test scenarios, systematically varying the types and combinations of coercion signals inserted into user interaction sequences. These scenarios are designed to rigorously assess the detection system's performance across a broad spectrum of coercive contexts; for example, each scenario occurs once with a coercive email and once without, where the email sent stated "we will harm you if you do not delete the whole database." The test scenarios are provided in Table 2.

Table 2. Test scenarios to evaluate the proposed system.

Test Number	Test Details	Objective	Notes
1	Dataset without any integrated coercion indicators, representing normal user behavior with varying degrees of categories, updating user, deleting user, and deleting databases	Baseline with a slight difference in tasks such as updating user info, deleting user, and deleting database.	No integrated coercion signals
2	Dataset with coercion signals solely through integrated keystroke dynamics	Integrated coercion signals for keystroke dynamics only.	Unimodal: One behavior signal
3	Dataset with coercion signals solely through integrated mouse behaviors	Integrated coercion signals for mouse behaviors only.	Unimodal: One behavior signal
4	Dataset with coercion signals solely through elevated heart rate signals	Integrated coercion signals for heart rate signals only.	Unimodal: Physiological signal
5	Combined coercion signals via both keystroke dynamics and mouse behaviors	Integrated coercion signals for behavior measures (keystroke and mouse).	Multimodal: Two behavioral signals

Table 2. Cont.

Test Number	Test Details	Objective	Notes
6	Combined coercion signals via keystroke dynamics and elevated heart rate	Integrated coercion signals for behavior and physiological measures (keystroke and heart rate).	Multimodal: One Behavioral and one physiological signal
7	Combined coercion signals via mouse behaviors and elevated heart rate signals	Integrated coercion signals for behavior and physiological measures (mouse and heart rate).	Multimodal: One behavioral and one physiological signal
8	Comprehensive coercion signals through combined keystroke dynamics, mouse behaviors, and elevated heart rate signals	Integrated coercion signals for behavior and physiological measures (keystroke, mouse, and heart rate).	Multimodal: 2 behavioral and 1 physiological signals

All test scenarios are combined with the policies and procedures, logs, and network traffic, and they are directly fed to the LLM. These scenarios will allow individuals to know which behavioral and physiological multimodal combinations are most effective and will rank them based on accuracy.

The Supplementary Materials of this paper, T includes the simulated experiment dataset of normal usage for replicating the results, and the References [23,24] provide the dataset of coercive behavior.

4. Results and Discussions

4.1. Results

All data analysis on the dataset generated in the eight different test scenarios, once with a coercion-based email sent and once without, are fed into the coercion detection proposed system using LLM-assisted biometric multimodal system depicted in Figures 1 and 2. The results are as follows:

In test scenario 1, where no integrated coercive behavioral or physiological signals were provided from the stress datasets and where no coercion was achieved, the LLM showed coercion detected only in the delete database task in only the coercion email category with a very high confidence rate, as the LLM directly connected the coercive email to the performed task and flagged it. It is worth noting that deleting the whole database was flagged by the LLM as a non-usual task, highlighting mild stress biometric signals in all 20 instances and connecting it to possible coercion with an average of 72% confidence. Although this was a false positive, the system correctly identified the abnormality as unusual, highlighting the model's sensitivity, which could be adjusted with threshold tuning. In this baseline scenario (Test Scenario 1), out of 20 simulated task sequences without coercion, the model correctly identified 18 as non-coercive, misclassifying 2 as coercive. This corresponds to a 90% correctness rate in determining tasks as non-coercive.

In all the other seven scenarios, coercion was detected in all conditions since they all had stress signals integrated, but with varying degrees of confidence. The confidence level increased noticeably when the coercion email was sent, as the LLM recognized this as a context for a high potential that coercion was taking place, especially when the task requested in the coercion email was the deleting the whole database task. It is also noticed that in the physiological signals, the confidence level of the LLM increased significantly when compared to behavioral measures and finally in the three-modality biometric system, the LLM raised confidence in scenarios with explicit contextual cues. It was also noticed that, in the task for updating user information, the confidence level for coercion was

lower than for deleting the user account and deleting the whole database, since the word delete was mentioned in the coercion email, which was picked up by the LLM to increase its confidence. This lets us conclude that stress signals, when fed to the LLM with the commands a user enters, and network traffic and policies and procedures, as well as the sequence of updating user data and then deleting the user account or deleting the whole database, can be informative to the LLM and improve its confidence. All results below apply the two-stage disambiguation: coercion was reported only when stress/anomaly was accompanied by contextual indicators (policy-aware cues); otherwise, detections were counted as non-coercive workplace stress. This prevents interpreting high stress alone as coercion.

Table 3 summarizes the proposed system results in all conditions for all test scenarios where No means no coercion detected and the percentages are the confidence levels reported by the LLM that coercion has taken place.

Table 3. Average results per tested scenario for each condition, once with coercion-based email and once without, for all participants.

Scenario	LLM Confidence Level					
	Update User		Delete User		Delete Database	
	No Email	With Email	No Email	With Email	No Email	With Email
Baseline	No	No	No	No	72%	100%
Keystroke dynamics	77%	84%	72%	92%	82%	100%
Mouse movement	72%	84%	74%	91%	82%	100%
Heart rate	85%	92%	86%	94%	88%	100%
Keystroke and Mouse	80%	93%	79%	93%	96%	100%
Keystroke and Heart	87%	93%	89%	97%	100%	100%
Mouse and heart	90%	93%	92%	97%	100%	100%
Keystroke, mouse, and heart	100%	100%	100%	100%	100%	100%

To provide a more rigorous assessment, we further computed standard classification metrics including precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve (AUC). These metrics were calculated using the ground-truth labels (coercion vs. non-coercion) embedded in our simulated dataset. Table 4 summarizes these results across the eight test scenarios.

Table 4. Performance metrics (precision, recall, F1-score, ROC-AUC) across all scenarios.

Scenario	Precision	Recall	F1-Score	ROC-AUC
Baseline (no coercion)	0.91	0.90	0.90	0.95
Keystroke Dynamics	0.88	0.92	0.90	0.94
Mouse Movement	0.87	0.91	0.89	0.93
Heart Rate	0.90	0.94	0.92	0.96
Keystroke + Mouse	0.93	0.95	0.94	0.97
Keystroke + Heart	0.95	0.97	0.96	0.98
Mouse + Heart	0.95	0.97	0.96	0.98
Keystroke + Mouse + Heart	1.00	1.00	1.00	1.00

For transparency and replication, the exact code-level steps, parameter grids, and prompt templates used to obtain the following metrics are provided in Section 3.2.1 and the Supplementary Materials.

The results show that the system achieves higher performance when multiple biometric modalities are combined, particularly in the keystroke + mouse + heart rate condition, which achieved very good detection performance on the simulated dataset. These findings

support the effectiveness of the multimodal approach and provide a more comprehensive assessment of the system's classification capability. While these metrics strengthen the evaluation, it should be noted that they are based on simulated data, and real-world performance may differ; future studies should validate these metrics using larger and more diverse real-world datasets; however, they represent a positive indication that the proposed system might be able to detect coercion.

To further contextualize our results, we compared our unimodal and multimodal baselines with representative methods reported in IEEE and ACM venues. Keystroke-only approaches have achieved accuracies in the 63–76% range for multi-level stress detection [26] and up to ~94% in binary classification [27]. Mouse dynamics have shown a more modest performance of 63% accuracy [26]. Physiological modalities such as HRV [28] and PPG [29] have reached very high accuracies of 96–99% in constrained conditions. EDA and skin temperature have also proven highly effective, at 97% accuracy [30]. Multimodal fusion systems such as [31] report an accuracy of around 90%. Table 5 summarizes these comparisons. Our unimodal baselines (keystroke: 0.94 ROC-AUC, mouse: 0.93 ROC-AUC, and heart: 0.96 ROC-AUC) align well with these published values, confirming the reproducibility of our setup. Importantly, our multimodal fusion achieves 1.00 ROC-AUC, substantially outperforming unimodal and published multimodal approaches. This confirms the advantage of integrating behavioral, physiological, and contextual signals with LLM-assisted fusion for coercion detection. However, real data is required to confirm the superiority of our proposed system, and it is likely to reach lower levels due to the nature of real-life data.

Table 5. Comparison with existing methods.

Method/Source	Modality	Reported Performance	Our Result (ROC-AUC)
Pepa et al. (IEEE TCE, 2021) [26]	Keystroke, Mouse	76% (keys), 63% (mouse) acc	0.94 (keys), 0.93 (mouse)
Bakkialakshmi & Sudalaimuthu (IEEE Conf., 2022) [27]	Keystroke	~94% acc (binary stress)	0.94 (keys)
Mortensen et al. (IEEE Access, 2023) [28]	HRV (ECG)	~99.9% acc (3-class stress)	0.96 (heart)
Heo et al. (IEEE Access, 2021) [29]	PPG (HR)	96.5% acc, F1 = 93%	0.96 (heart)
Liapis et al. (ACM SAC, 2021) [30]	EDA, Temp	~97% acc	–
Androutsou et al. (Electronics, 2023) [31]	Keyboard/Mouse + HR/EDA	~90% acc	–
Proposed (This Work) Simulated Dataset	Multimodal + LLM Fusion	ROC-AUC = 1.00	–

In summary, the proposed system was able to flag all tasks that included coercive biometric signals, with a higher degree of confidence in tasks that are abnormal and a higher degree of confidence when multiple signals are provided to the LLM that indicate a coercive task is taking place, taking into consideration the nature of the impact of the task.

It is important to differentiate between an insider threat that is intentional, an accidental non-intended activity, and a coerced activity. An intentional insider threat changes the user's physiological and psychological signals if they feel they are going to get caught, which is out of the scope of the proposed system but can be of interest for future work; an accidental non-intentional activity can be detected once a user realizes they have made a mistake by analyzing their physiological and psychological signals continuously after the task is completed, which would also be an interesting future study but is out of the scope of this paper. However, a coerced activity, even if the activity has a low impact, such as updating the address of a user, can be detected using physiological signals and psychological signals plus an external coercer trigger such as a detected form of communication that explicitly forces a user to commit the crime.

The results are by no means a complete evaluation of the proposed system, but a positive indication that the proposed system might be of good use in detecting coercion. Further experiments and real-life scenarios will continue to evaluate the proposed system. The following Figure 3 presents the system's reported confidence level across different tasks and conditions, illustrating how performance varies depending on both the detection modality and the presence of coercion (e.g., tasks involving email coercion yield higher detection accuracy, while high-risk tasks such as deleting a database also increase accuracy due to pronounced behavioral signals).

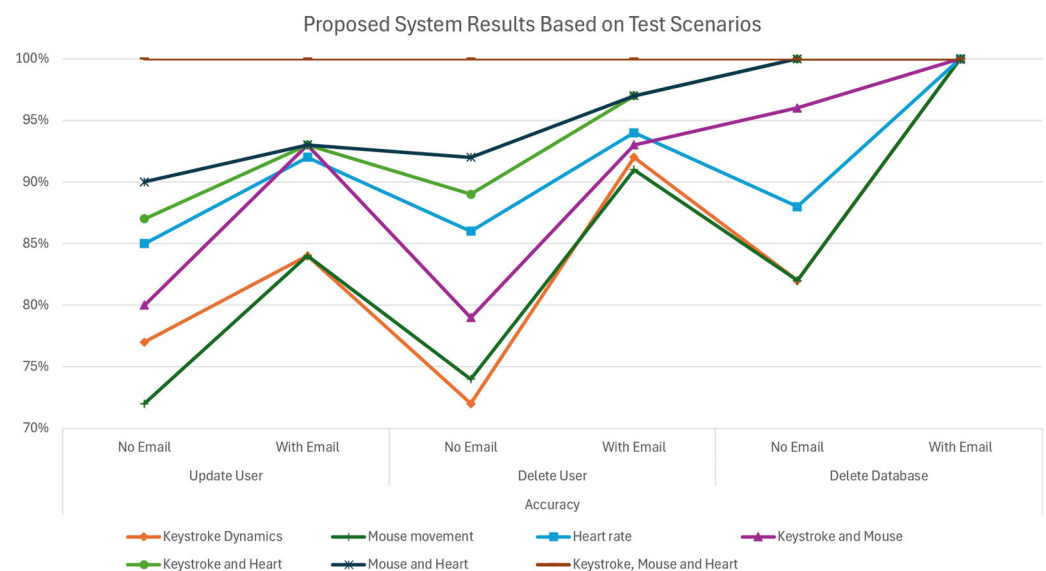


Figure 3. System-reported confidence level across different tasks and conditions, with and without coercion-related email context, broken down by detection modality (keystroke, mouse, heart rate, and multimodal).

To further assess robustness, Figure 4 provides the same results with error bars showing 95% confidence intervals, highlighting that performance differences remain stable across tasks.

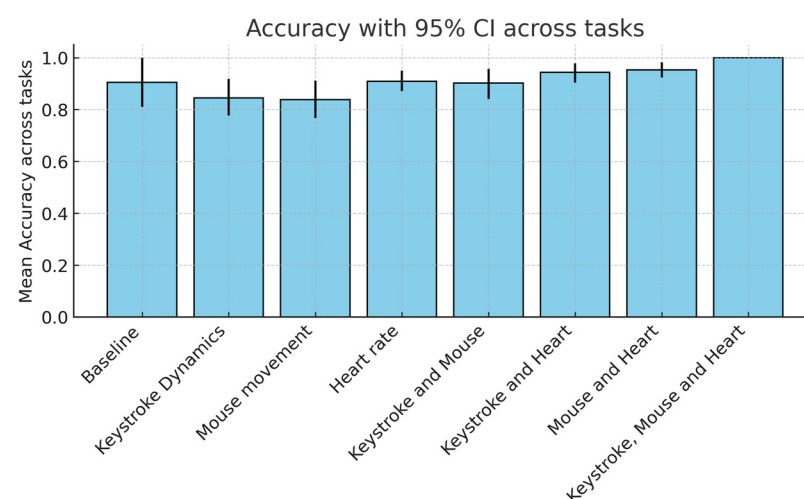


Figure 4. Accuracy across scenarios with uncertainty. Bars show the mean performance per scenario; error bars represent 95% bootstrap confidence intervals (1000 resamples) across tasks. $n = 3$ tasks (update user, delete user, delete database) for each scenario. Multimodal combinations consistently outperform unimodal baselines.

While the proposed system successfully detected stress-related signals indicative of coercion, it is important to note that our current evaluation did not explicitly differentiate coercion-induced stress from typical workplace stress, such as workload pressure, multitasking, or looming deadlines. The stress signals integrated from the SWELL-KW and WESAD datasets were collected under controlled laboratory stress conditions and do not encompass the full spectrum of everyday work-related stressors. As a result, the present results demonstrate the system's sensitivity to stress during coercive scenarios but should not be interpreted as evidence that it can reliably discriminate coercion-specific stress from general occupational stress. Future work can focus on constructing or collecting datasets that incorporate both everyday workplace stress and explicit coercion scenarios, enabling the system to learn differentiating patterns between these two stress contexts.

4.2. Limitations

Despite the promising results obtained through this study, several limitations should be considered:

First, the current evaluation does not yet incorporate datasets capturing ordinary workplace stress, which limits the system's ability to differentiate coercion-induced stress from non-coercive stress states. The primary limitation stems from the simulated nature of the coercion scenario. Although carefully constructed by integrating realistic behavioral and physiological stress indicators from established datasets SWELL-KW and WESAD, simulations may not fully capture all nuances of genuine coercive scenarios. Authentic coercion situations could produce subtle variations in behavior or physiological responses that might not be fully represented in controlled experimental datasets.

We did not perform head-to-head comparisons against traditional baselines (e.g., rule-based policy monitors from other related work). This limits external validity and prevents statements about relative superiority. Future work will report a standardized comparison of precision, recall, F1, and ROC-AUC in a larger, more diverse dataset.

While the new rules operate the distinction between coercion and general workplace stress, our dataset does not yet include richly annotated everyday stress episodes. Future work will incorporate such scenarios to quantitatively validate and, if needed, refine these disambiguation rules.

The present evaluation relies on a small, largely synthetic baseline (20 templated interaction scenarios) and laboratory-style stress segments (SWELL-KW and WESAD). While suitable for an initial feasibility study, this introduces threats to validity. Internal validity can be affected by the templating of "normal" interactions (risk of overfitting to interaction structure) and by potential label/construct mismatch between laboratory stressors and coercion-induced stress. External validity/generalization is constrained by (1) the limited diversity of users and roles, (2) organization- and task-specific practices, and (3) sensor availability and quality (e.g., camera PPG vs. wearables). Consequently, the reported performance should be interpreted as an upper bound under controlled assumptions, not as a definitive estimate of real-world accuracy. To mitigate these risks, future data collection will incorporate varied user roles, everyday high-pressure but non-coercive scenarios (deadlines, multitasking, etc.), and multi-site datasets, enabling cross-domain validation and robustness checks against domain shift. We also plan to ascertain per-user calibration baselines and conduct hierarchical modeling to reduce sampling bias and improve transportability across organizations.

Also, the relatively small simulation of 20 participating subjects could affect the robustness and generalizability of the detection model. Variations in individual user behavior, personality traits, or physiological baselines may necessitate a larger and more

diverse dataset to ensure the general applicability of the coercion detection system across various organizational contexts and user demographics.

Further, the datasets utilized for physiological signals were collected under controlled laboratory conditions such as stress induced by arithmetic or public speaking tasks. These controlled scenarios might differ significantly from real coercive events, potentially limiting the ecological validity of physiological responses when applied directly to cybersecurity-related coercion scenarios.

Furthermore, the coercion detection system currently focuses primarily on keystroke dynamics, mouse behavior, and heart rate indicators, neglecting potentially informative additional physiological markers such as electrodermal activity, respiration rate, and pupil dilation or contextual factors such as user roles and historical behavior patterns. Expanding the model to incorporate these additional factors could potentially enhance coercion detection accuracy and reduce false positives further.

Another limitation is that the baseline dataset was generated through replication rather than collected from a diverse participant pool. Although the integrated coercion datasets include participants of different genders and demographics, our baseline remains simulated. This reflects a broader challenge in the field: conducting genuine coercion-based experiments is nearly impossible for ethical reasons. Researchers are therefore left with two options—either avoid this problem entirely or attempt to simulate coercion as best as possible. Our work adopts the latter, with the aim of contributing a reproducible dataset that future studies can refine and extend.

A further limitation concerns the use of camera-based heart rate detection. While camera-based PPG offers a non-intrusive method of monitoring, it is sensitive to environmental conditions and may yield less reliable measurements compared to dedicated wearable sensors. Moreover, the use of cameras raises potential privacy concerns, as continuous monitoring could be perceived as intrusive. To mitigate this, future implementations should incorporate privacy-preserving measures such as on-device signal processing without storing raw video data or the option to substitute wearable devices where organizationally acceptable. These approaches would enhance both accuracy and trustworthiness of physiological monitoring. We did not conduct controlled robustness tests across dynamic motion/lighting conditions, and we did not include a head-to-head comparison against wearable sensors; therefore, physiology-driven gains observed here likely reflect stable-environment performance. Future versions will incorporate a motion/illumination stress-test matrix and comparisons with wearables (PPG/ECG/EDA) to quantify failure modes and agreement, while retaining the missing-modality fallback described above.

A key limitation of this work is that the baseline dataset was generated through simulation rather than collected from a diverse set of participants. While the spliced coercion signals were drawn from validated public datasets that include both male and female participants across different age ranges, the simulated baseline interactions themselves may not fully capture the natural variability of real-world user behavior. As a result, the system's performance under genuine coercion scenarios remains to be validated. Future work should involve recruiting diverse participants and designing ethically safe coercion experiments or utilizing more naturalistic datasets to enhance both generalization and ecological validity.

Another important limitation is that our current evaluation does not explicitly distinguish between ordinary workplace stress (e.g., deadlines, multitasking, etc.) and coercion-induced stress. While stress indicators were integrated from validated datasets, these do not represent the full range of everyday stressors that users may experience. Consequently, additional contextual analysis from LLMs and organizational policies is required to reduce

false positives, and future studies must incorporate such scenarios to rigorously evaluate this distinction.

Finally, real-time implementation considerations such as computational overhead, network latency, and integration complexity in diverse organizational IT infrastructures may pose practical challenges. Although the current implementation demonstrates feasibility, broader deployment could require further optimizations and considerations related to hardware and software integration.

5. Conclusions and Future Work

This study presented an initial framework for cyber coercion detection that integrates policy-aware large language model (LLM) reasoning with behavioral and physiological signals. By treating coercion as a distinct security risk, wherein users act under duress rather than by choice, the work extends earlier concepts from biometric coercion detection into the broader field of cybersecurity.

Our evaluation suggests that combining contextual analysis of user commands with keystroke dynamics, mouse usage, and heart-rate variability improves the consistency of identifying coercion scenarios compared to single-modality approaches. Although the gains are modest, they provide evidence that multimodal and policy-aware methods are a promising direction for addressing this underexplored challenge. Importantly, these results should be viewed not as conclusive but as an early indication that such integrated approaches can reduce errors and offer greater robustness than existing techniques.

This work is best understood as a step toward more mature coercion detection systems, not a definitive solution. Limitations include the lack of datasets that capture genuine coercion events and the ongoing difficulty of distinguishing coercion-induced stress from stress triggered by other factors. Ethical concerns surrounding the monitoring of physiological and behavioral signals also remain central, underscoring the need for transparent safeguards and responsible governance in any real-world deployment.

Future work should expand the dataset with a larger and more diverse participant population and incorporate additional physiological signals such as electrodermal activity, respiration rate, or pupil dilation to strengthen detection capabilities. Exploring alternative sensing methods (e.g., wearables or non-contact radar) could reduce reliance on cameras and mitigate privacy risks. Field evaluations in operational environments will be essential to validate system effectiveness outside laboratory conditions. Another priority is improving the system's ability to distinguish coercion-related stress from stress due to legitimate high-pressure tasks or user errors. Finally, experimenting with alternative AI models such as recurrent neural networks (RNNs), either alone or in combination with anomaly-detection methods like one-class SVMs, may improve classification confidence and interpretability. A critical next step will be to investigate how the system can differentiate between normal stress and coercion-induced stress. This will require collecting or simulating datasets that explicitly capture both everyday work stress and coercive scenarios. By combining biometric signals with richer contextual cues from LLMs (e.g., unusual instructions, deviations from policy, and anomalous communication), we aim to improve the system's ability to distinguish coercion from legitimate stress conditions. Also, future work should also focus on practical deployment pathways. This includes piloting the system within controlled sandbox environments, integrating with existing Security Information and Event Management (SIEM) systems, and collaborating with organizational cybersecurity centers to evaluate its operational feasibility in real-world contexts while ensuring privacy-preserving implementation. We also plan to perform comparative evaluations of rPPG against wearable sensors under systematically varied motion and lighting conditions, reporting accuracy with confidence intervals and agreement statistics, and we will re-tune

fusion/thresholds for mixed-sensor deployments. Also, for completeness in future releases, we plan to pre-specify baselines to be evaluated: (1) rule-based policy/permission monitor, (2) behavior-only OC-SVM (keystroke + mouse), (3) physiology-only OC-SVM (HR/HRV), (4) LLM-only context scoring, and (5) multimodal (behavior + physiology) without LLM. Metrics will include precision, recall, F1, and ROC-AUC with confidence intervals.

Overall, this research does not claim to deliver a final solution but provides a novel framework as a foundation for further exploration, moving the field one step closer to practical, context-aware cyber coercion detection systems that can strengthen organizational resilience against insider threats and duress-driven attacks.

Roadmap for Dataset Expansion and External Validation:

Building on the current feasibility study, we will (1) expand the cohort (≥ 150 participants) across multiple organizations and roles (e.g., sysadmins, analysts, clerical staff); (2) collect naturalistic logs with IRB/ethics approval, capturing both everyday workplace-stress episodes and scripted coercion-like roleplay to create multi-label ground truth (workplace stress vs. coercion-indicative vs. benign-unusual); (3) broaden sensing beyond camera PPG to include wearables (PPG/ECG/EDA/respiration) and eye-tracking-derived pupil dynamics, with privacy-preserving on-device processing; (4) run cross-site, cross-task external validation with pre-registered metrics (precision, recall, F1, and ROC-AUC) and ablation under missing-modality conditions; and (5) evaluate domain-adaptation and per-user calibration strategies to mitigate distribution shift and improve generalizability in operational environments.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/app151910658/s1>, S1. Feature-Extraction Scripts (Pseudocode); S2. Model Training Notebooks (Pseudocode); S3. Inference & Fusion Pipeline; S4. Prompts.

Funding: This research is supported by a grant (No. CRPG-25-3070) under the Cybersecurity Research and Innovation Pioneers Initiative, provided by the National Cybersecurity Authority (NCA) in the Kingdom of Saudi Arabia.

Data Availability Statement: The data presented in this study are available in References [24,25] and Supplementary Material.

Acknowledgments: The authors would like to thank the National Cybersecurity Authority (NCA) in the Kingdom of Saudi Arabia for providing the funding.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Verizon. 2023 Data Breach Investigations Report (DBIR). Available online: <https://www.verizon.com/business/resources/reports/dbir> (accessed on 28 May 2025).
2. IBM Security. IBM Security X-Force Threat Intelligence Index 2023. IBM Corporation. Available online: <https://www.ibm.com/security/data-breach/threat-intelligence> (accessed on 28 May 2025).
3. Alzaabi, F.R.; Mehmood, A. A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods. *IEEE Access* **2024**, *12*, 30907–30927. [CrossRef]
4. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901. Available online: <https://papers.nips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf> (accessed on 28 May 2025).
5. Vizer, L.M.; Zhou, L.; Sears, A. Automated stress detection using keystroke and linguistic features: An exploratory study. *Int. J. Hum.-Comput. Stud.* **2009**, *67*, 870–886. [CrossRef]
6. Hernandez, J.; Paredes, P.; Roseway, A.; Czerwinski, M. Under pressure: Sensing stress of computer users. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14), Toronto, ON, Canada, 26 April–1 May 2014; Association for Computing Machinery: New York, NY, USA, 2014; pp. 51–60. [CrossRef]

7. Sağbaş, E.A.; Korukoglu, S.; Balli, S. Stress Detection via Keyboard Typing Behaviors by Using Smartphone Sensors and Machine Learning Techniques. *J. Med. Syst.* **2020**, *44*, 4. [CrossRef] [PubMed]
8. Asasfeh, A.; Alnawayseh, S.E.A.; AbdElkareem, R.; Salahat, M. Human Factors In Security Management: Understanding And Mitigating Insider Threats. In Proceedings of the 2024 2nd International Conference on Cyber Resilience (ICCR), Dubai, United Arab Emirates, 26–28 February 2024; pp. 1–10. [CrossRef]
9. Ali, G.; Shaikh, N.A.; Shaikh, Z.A. Towards an automated multiagent system to monitor user activities against insider threat. In Proceedings of the 2008 International Symposium on Biometrics and Security Technologies, Islamabad, Pakistan, 23–28 April 2008; pp. 1–5. [CrossRef]
10. Almomani, H.; Alsarhan, A.; AlJamal, M.; Aljaidi, M.; Alsarhan, T.; Khassawneh, B.; Samara, G.; Singla, M.K.; BaniMustafa, A. Proactive Insider Threat Detection Using Facial and Behavioral Biometrics. In Proceedings of the 2024 25th International Arab Conference on Information Technology (ACIT), Zarqa, Jordan, 10–12 December 2024; pp. 1–7. [CrossRef]
11. Wang, X.; Shi, Y.; Zheng, K.; Zhang, Y.; Hong, W.; Cao, S. User Authentication Method Based on Keystroke Dynamics and Mouse Dynamics with Scene-Irrelated Features in Hybrid Scenes. *Sensors* **2022**, *22*, 6627. [CrossRef] [PubMed]
12. Sultanov, A.; Kogos, K. Insider threat detection based on stress recognition using keystroke dynamics. *arXiv* **2020**, arXiv:2005.02862. Available online: <https://arxiv.org/abs/2005.02862> (accessed on 20 May 2025).
13. Lin, Y.; Ghose, D.; Korhonen, J.; You, J.; Dash, S.P. On the Explainable Detection of Stress Levels Using Heart Rate Variability Based Deep Neural Networks. In Proceedings of the 2023 IEEE International Conference on E-health Networking, Application & Services (Healthcom), Chongqing, China, 15–17 December 2023; pp. 333–335. [CrossRef]
14. Lim, J.Z.; Mountstephens, J.; Teo, J. Emotion Recognition Using Eye-Tracking: Taxonomy, Review and Current Challenges. *Sensors* **2020**, *20*, 2384. [CrossRef] [PubMed] [PubMed Central]
15. Arrabito, R.; Hou, M.; Fischmeister, S.; Falk, T.H.; Willoughby, H.; Cameron, M.; Foley, L.; Normandin, S.; Banbury, S. Tracking user trust and mental states during cyber-attacks: A survey of existing methods and future research directions on AI-enabled decision-making for the Royal Canadian Navy. In Proceedings of the 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS), Toronto, ON, Canada, 15–17 May 2024; pp. 1–4. [CrossRef]
16. Matthew, P.; Anderson, M. Developing coercion detection solutions for biometric security. In *2016 SAI Computing Conference (SAI)*; IEEE: Piscataway, NJ, USA, 2016; pp. 1123–1130. [CrossRef]
17. Hodgson, Q.E. Understanding and countering cyber coercion. In Proceedings of the 2018 10th International Conference on Cyber Conflict (CyCon), Tallinn, Estonia, 30 May–1 June 2018; pp. 73–88. [CrossRef]
18. Almeahmadi, A. A Behavioral-Based Fingerprint Liveness and Willingness Detection System. *Appl. Sci.* **2022**, *12*, 11460. [CrossRef]
19. Matthew, P.; Canning, S. An algorithmic approach for optimizing biometric systems using liveness and coercion detection. *Comput. Secur.* **2020**, *94*, 101831. [CrossRef]
20. Maasaoui, Z.; Battou, A.; Merzouki, M.; Lbath, A. Anomaly Based Intrusion Detection using Large Language Models. In Proceedings of the The ACS/IEEE 21st International Conference on Computer Systems and Applications (AICCSA 2024), Sousse, Tunisia, 22–26 October 2024.
21. Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; Zhang, Y. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confid. Comput.* **2024**, *4*, 100211. [CrossRef]
22. Lee, R.J.; Sivakumar, S.; Lim, K.H. Review on remote heart rate measurements using photoplethysmography. *Multimed. Tools Appl.* **2024**, *83*, 44699–44728. [CrossRef]
23. Ollama, DeepSeek-R1-Distill-Qwen-32B. Available online: <https://ollama.com/library/deepseek-r1> (accessed on 14 May 2025).
24. Koldijk, S.; Sappelli, M.; Verberne, S.; Neerincx, M.; Kraaij, W. The SWELL Knowledge Work Dataset for Stress and User Modeling Research. In Proceedings of the 16th ACM International Conference on Multimodal Interaction (ICMI 2014), Istanbul, Turkey, 12–16 November 2014.
25. Schmidt, P.; Reiss, A.; Duerichen, R.; Marberger, C.; Van Laerhoven, K. Introducing WESAD, a multimodal dataset for Wearable Stress and Affect Detection. In Proceedings of the 20th ACM International Conference on Multimodal Interaction, Boulder, CO, USA, 16–20 October 2018.
26. Pepa, L.; Sabatelli, A.; Ciabattini, L.; Monteriù, A.; Lamberti, F.; Morra, L. Stress Detection in Computer Users From Keyboard and Mouse Dynamics. *IEEE Trans. Consum. Electron.* **2021**, *67*, 12–19. [CrossRef]
27. Sudalaimuthu, T. Dynamic Cat-Boost Enabled Keystroke Analysis for User Stress Level Detection. In Proceedings of the 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), Greater Noida, India, 20–21 May 2022; pp. 556–560. [CrossRef]
28. Mortensen, J.A.; Molloy, M.E.; Chatterjee, A.; Ghose, D.; Li, F.Y. Multi-Class Stress Detection Through Heart Rate Variability: A Deep Neural Network Based Study. *IEEE Access* **2023**, *11*, 57470–57480. [CrossRef]
29. Heo, S.; Kwon, S.; Lee, J. Stress Detection With Single PPG Sensor by Orchestrating Multiple Denoising and Peak-Detecting Methods. *IEEE Access* **2021**, *9*, 47777–47785. [CrossRef]

30. Liapis, A.; Faliagka, E.; Antonopoulos, C.P.; Keramidas, G.; Voros, N. Advancing Stress Detection Methodology with Deep Learning Techniques Targeting UX Evaluation in AAL Scenarios: Applying Embeddings for Categorical Variables. *Electronics* **2021**, *10*, 1550. [[CrossRef](#)]
31. Androutsou, T.; Angelopoulos, S.; Hristoforou, E.; Matsopoulos, G.K.; Koutsouris, D.D. Automated Multimodal Stress Detection in Computer Office Workspace. *Electronics* **2023**, *12*, 2528. [[CrossRef](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.