Forensic Science International: Digital Investigation 53 (2025) 301931

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS USA 2025 - Selected Papers from the 25th Annual Digital Forensics Research Conference USA

Your forensic AI-assistant, SERENA: Systematic extraction and reconstruction for enhanced A2P message forensics

Jieon Kim, Byeongchan Jeong, Seungeun Park, Sangjin Lee, Jungheum Park *

School of Cybersecurity, Korea University, Seoul, South Korea

A R T I C L E I N F O

Keywords:
Digital forensics
Application-to-person (A2P) message
Textual data processing
AI-assisted data analysis
Named entity recognition (NER)
Large language mode (LLM)
Forensic tool development

A B S T R A C T

The integration of physical and online activities in today's hyper-connected world has blurred previously distinct boundaries. Online actions such as reservations, payments, and logins generate application-to-person (A2P) messages, which serve as valuable datasets for tracking user behavior. Although A2P messages from different service providers may vary in structure, the information within each message can be systematically normalized based on user behavior and service characteristics. However, traditional forensic tools have been unable to effectively identify and extract such forensically valuable information from these A2P messages. In this study, we leverage large language models (LLMs) combined with prompt engineering to analyze A2P messages from multiple service providers, addressing the limitations of existing forensic tools in extracting meaningful insights from unstructured or semi-structured text stored in messages and emails. The proposed methodology employs A2P messages to elaborately reconstruct user activity, enabling digital forensic investigations to identify case-relevant information with enhanced efficiency and accuracy.

1. Introduction

In the modern era, the lines between physical and online activities have become increasingly blurred, creating a seamless integration of behaviors across digital and real-world domains. Automated service messages, triggered by online actions such as reservations, payments, lodging, ticket purchases, and logins, have emerged as a valuable source of data for understanding user behavior (Kim et al., 2023). In this context, large volumes of application-to-person (A2P) messages are being generated and stored on users' local devices as well as remote cloud servers. However, traditional forensic tools struggle to extract meaningful insights from these A2P messages, leading to inefficiencies for digital forensic practitioners.

This study proposes a novel methodology that leverages large language models (LLMs) with prompt engineering to address this gap. By analyzing A2P messages generated by multiple different service providers, the proposed approach automatically extracts meaningful forensic insights, enabling investigators to reconstruct user behaviors with greater accuracy and efficiency. Given that understanding a suspect's past activities is crucial in digital forensics, an AI-driven approach can significantly enhance investigative workflows.

1.1. Motivation

Digital forensics plays a vital role in modern investigations, often relying on vast datasets collected from diverse digital sources. While A2P messages provide structured and reliable information about user activities, the sheer volume of such data presents significant challenges for investigators. Traditional forensic tools focus on data collection rather than insight extraction, which can result in missed connections and overlooked evidence.

Currently, most open source, free and commercial digital forensic analysis tools support keyword search capabilities for raw data or parsed text related to email and messaging services. However, to utilize keyword search capabilities for analyzing A2P messages, investigators must manually compile a list of search terms after identifying case-relevant information. To address this challenge, AI-assisted forensic tools are emerging to align with recent advancements. For instance, Magnet AXIOM offers an AI-powered feature that automatically categorizes messages based on pre-defined topics (Magnet Forensics, 2018). While a notable effort, this feature remains quite limited in scope, as it is primarily designed to detect person-to-person (P2P) messages related to sexually explicit content, luring and kidnapping. Consequently, current

* Corresponding author.
E-mail addresses: kijie@korea.ac.kr (J. Kim), naaya@korea.ac.kr (B. Jeong), pse0103@korea.ac.kr (S. Park), sangjin@korea.ac.kr (S. Lee), jungheumpark@korea.ac.kr (J. Park).

https://doi.org/10.1016/j.fsidi.2025.301931

forensic tools struggle to effectively analyze A2P messages, which often contain crucial forensic evidence such as transaction records, login notifications, and travel confirmations.

Given the increasing reliance on online services and digital transactions, A2P messages have become a vital forensic resource. These messages offer timestamped records of user interactions, providing valuable insights into activity patterns, account access, and financial behaviors. However, effectively utilizing this data requires an automated, scalable approach capable of analyzing text data, which varies in structure and format across different services.

Large Language Models (LLMs) offer a transformative solution to this challenge (Cho et al., 2024). Unlike rule-based forensic tools, LLMs excel at natural language understanding, making them particularly suited for analyzing A2P messages that contain context-rich data with forensically valuable named entities such as timestamps, names, locations and more. By employing advanced prompt engineering, LLMs can interpret contextual information, extract key forensic insights, and detect patterns that might otherwise go unnoticed. This approach enhances investigative efficiency, reduces cognitive load on forensic practitioners, and ensures a higher degree of accuracy in identifying and analyzing user behaviors (K et al., 2023).

1.2. Contribution

This study provides the following contributions to the digital forensic community:

- We propose the use of LLMs with prompt engineering to analyze A2P messages, leveraging their semi-structured nature to extract actionable insights efficiently.
- The proposed methodology bridges the gap between data collection and interpretation, enabling investigators to identify relevant user activities and reconstruct timelines with greater precision.
- By developing a proof-of-concept tool that utilizes LLMs in a forensic context, this study demonstrates their effectiveness in analyzing A2P messages, which contain text data with varying structures and formats across different services.

The remainder of this paper is structured as follows. **Section** 2 describes what A2P messages are and their characteristics. Additionally, it examines related studies. **Section** 3 outlines our proposed methodology for classifying A2P messages and extracting named entities from the classified messages. **Section** 4 introduces **SERENA**, a proof-of-concept tool developed based on the proposed methodology. **Section** 5 presents the design and results of experiments to evaluate the performance and effectiveness of **SERENA**. Finally, **Section** 6 summarizes our contributions and provides suggestions for future research directions.

2. Background and related work

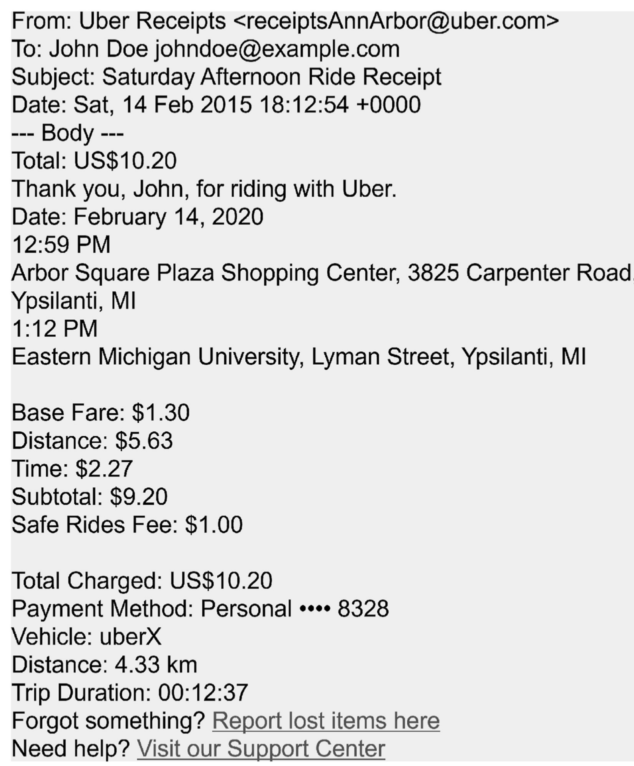
2.1. A2P messages

Unlike a person-to-person (P2P) message, an application-to-person (A2P) message is automatically sent to users by a service provider for various purposes (Twilio Inc., 2024). These messages are delivered through multiple channels, including social networking service (SNS) apps, text messages, and emails, ensuring that users receive important notifications across different platforms. A2P messages are typically triggered when users perform specific actions, such as registration, reservations, orders, deliveries, or billing updates. The behaviors associated with receiving A2P messages can be broadly categorized into several areas, including purchasing products, making reservations for services (e.g., restaurants, cinemas, accommodations), and using transportation. Table 1 presents different types of A2P messages along with sample messages.

The distinction between physical and online spaces is no longer clear,

Table 1
Types of A2P messages and real-world examples.

Order	
Service Name	**Example Message**
Amazon	Your Amazon.com Order of *ItemName*
Nike	Order Shipped (Nike.com #*OrderNumber*)
Transaction	
Service Name	**Example Message**
Finpay	Thank you, we have received your payment and will process your payment immediately
Cyber Payment	Here is the payment detail for your transaction on *ItemName* & *DateTime*
Booking	
Service Name	**Example Message**
Expedia	Expedia flight purchase confirmation
Virgin Australia	Retrieve your Virgin Australia Boarding Pass
Notification	
Service Name	**Example Message**
Emergency Alert	UPennAlert: All clear in the area of *location*
USPS	USPS-Click-N-Ship(R) Notification
Miscellaneous	
Service Name	**Example Message**
UNICEF	[UNICEF] Donation Registration Complete
University	Congratulations on your admission to *Name of University*


From: Uber Receipts <receiptsAnnArbor@uber.com>
To: John Doe johndoe@example.com
Subject: Saturday Afternoon Ride Receipt
Date: Sat, 14 Feb 2015 18:12:54 +0000
--- Body ---
Total: US$10.20
Thank you, John, for riding with Uber.
Date: February 14, 2020
12:59 PM
Arbor Square Plaza Shopping Center, 3825 Carpenter Road, Ypsilanti, MI
1:12 PM
Eastern Michigan University, Lyman Street, Ypsilanti, MI

Base Fare: $1.30
Distance: $5.63
Time: $2.27
Subtotal: $9.20
Safe Rides Fee: $1.00

Total Charged: US$10.20
Payment Method: Personal •••• 8328
Vehicle: uberX
Distance: 4.33 km
Trip Duration: 00:12:37
Forgot something? Report lost items here
Need help? Visit our Support Center

Fig. 1. Example of a well-organized and grammatically correct A2P message: This is an actual message (with some information redacted) emailed immediately after using a ride-sharing service. It contains various named entities useful for digital forensics, such as origin, destination, distance, duration, fare, timestamp, and more.

A2P messages can be regarded as a collective representation of user behaviors. Therefore, we observe that the past behavior could be predicted by analyzing the A2P messages sent from various services. As shown in Fig. 1, these A2P messages are clear, and typically use grammatically accurate language with minimal slang, making them highly suitable for analytical purposes. Furthermore, these characteristics make A2P messages well-suited for the application of large language models (LLMs).

2.2. Related work

2.2.1. Existing forensic tools and their limitations

In the modern era, people communicate through messages and emails, making their analysis essential in forensic investigations. Various forensic tools provide message and email analysis and visualization capabilities, and research in this area has been actively conducted.

Chen et al. (2017) proposed an email forensic system with enhanced visualization and correlation analysis. The system provides a keyword search, with results visualized through various charts and graphs. Notably, a tree network diagram represents sender–receiver relationships between email addresses. While effective for analyzing P2P messages, its applicability to A2P message analysis remains limited.

Ghafarian et al. (2020) evaluated various email forensic tools, which primarily rely on built-in search functions for filtering by name, email, and date. Aid4Mail and MailXaminer support regex-based extraction of URLs and phone numbers. Autopsy and OSForensics use predefined keyword lists for restricted content classification. While efficient, these tools remain domain-limited.

Pirzada et al. (2023) reviewed forensic analysis and visualization studies for mobile messaging apps. Most open-source, free and commercial forensic tools offer keyword searches but require case-specific knowledge for efficient data retrieval. They found that existing studies focused on P2P messages, often containing slang and typos, limiting their usefulness as digital forensic evidence.

2.2.2. Machine Learning and Natural Language Processing

To extract meaningful insights from large text datasets, Machine Learning (ML) and Natural Language Processing (NLP) have been widely applied across various domains. In forensic investigations, vast textual data must be analyzed efficiently. Existing studies have leveraged ML and NLP to process messages, emails, and logs, aiding faster, more accurate, and scalable evidence analysis.

Hina et al. (2021) proposed a system that classifies emails into four categories: normal, harassing, fraudulent, and suspicious. After pre-processing and feature extraction with TF-IDF, experiments were performed to classify emails using ML algorithms, such as Logistic Regression, Naive Bayes, and Random Forest. This could contribute to automating investigations by utilizing ML to process evidential text data.

Rai et al. (2023) performed sentiment analysis on user reviews from ride-sharing applications, classifying them into categories. The dataset was built by collecting reviews from apps like Uber and Lyft. Experiments using CNN, LSTM, and DistilBERT showed the pretrained DistilBERT model achieved the highest accuracy. This suggests pretrained transformer models can outperform other models in NLP tasks.

Tejaswini et al. (2024) developed an NLP-based deep learning model to detect depression from social media texts. Using fastText embeddings, CNN, and LSTM, the model analyzed Reddit and Twitter data, achieving higher accuracy than existing methods. The study highlights deep learning's effectiveness in extracting patterns and classifying depressive and non-depressive texts.

P2P messages often contain informal language, including slang and abbreviations, which makes them more challenging to process. However, to address this, Goel et al. (2024) proposed a smishing detection framework that incorporates advanced text normalization, using the NoSlang dictionary to replace informal language with formal equivalents. The study confirmed a significant accuracy improvement after standardization in classifying normal and smishing messages using Naive Bayes.

Both Shahbazi and Byun (2022) and Adkins et al. (2024) leverage NLP-based AI techniques to enhance digital forensic investigations by analyzing textual data. Shahbazi and Byun (2022) focuses on system security, applying NLP methods to detect threats and anomalies in online platforms. However, it lacks concrete examples or empirical results showing practical forensic applications. Meanwhile, Adkins et al. (2024) extracts key entities and relationships from digital communications, such as emails and social media, to identify individuals relevant to forensic cases; however, its findings have primarily been limited to analyzing user interests and have not yet been translated into practical forensic applications.

In summary, existing approaches using NLP and ML techniques have focused primarily on human-authored, domain-specific content such as P2P emails or social media posts, often relying on regular expressions or task-specific fine-tuning for entity extraction. In contrast, our approach is designed to handle semi-structured, machine-generated A2P messages originating from diverse services. Rather than retraining models or building service-specific parsers for each format, this study utilizes LLMs' contextual understanding to extract and normalize forensic entities across heterogeneous sources. This makes our method more scalable and practical for real-world forensic applications.

2.2.3. LLMs and digital forensics

To enhance the efficiency of investigative work, recent research in the field of digital forensics has increasingly focused on leveraging Large Language Models (LLMs) (Scanlon et al. (2023)).

Nguyen et al. (2023) fine-tuned a pretrained Llama 2 model using LoRA to detect online sexual predatory chat and abusive text. Experiments were conducted on English, Roman Urdu, and Urdu datasets, and the model performed excellently across all languages, highlighting its potential applicability to non-English languages as well. This suggests that fine-tuning a pretrained model, which has acquired general knowledge, can provide domain-specific expertise, making it effective in solving problems across various domains.

Both studies (Egersdoerfer et al. (2023); Chernyshev et al. (2024)) explore the application of large language models (LLMs) in forensic investigations by leveraging structured log data to enhance anomaly detection and evidence extraction. Egersdoerfer et al. (2023) examines ChatGPT for detecting anomalies in parallel file system logs (PFS logs), which record system activity in high-performance computing environments. By analyzing log patterns and summarizing behaviors, the approach improves security threat detection and reduces manual forensic effort. Similarly, Chernyshev et al. (2024) proposes LLM invocation logging to enhance forensic readiness in applications using LLMs. Analyzing invocation logs provides insights into user interactions, improving forensic efficiency. This LLM-based forensic analysis can extend to A2P message analysis, enhancing automation in digital investigations.

Michelet and Breitinger (2024) analyzed typical forensic reports to identify a general structure and then assessed the capabilities and limitations of LLMs (GPT-3.5 and Llama-2) in producing each part through a case study approach. Certain sections of a report follow a predefined structure or use specific terminology, making them well-suited for leveraging LLMs. In our study on A2P messaging, we also utilized this characteristic effectively.

Oh et al. (2024) proposed a prompt-based LLM called volGPT, which leverages the process information extracted using Volatility during the memory forensics process to triage suspicious processes. The experimental results using ransomware samples demonstrated high accuracy and efficiency in triage.

Park et al. (2024) suggested a hybrid classification model designed to extract critical information for investigative reports at both the token and sequence levels. The model, trained on a dataset constructed using GPT-3.5, demonstrated superior performance compared to baseline models in extracting key crime-related information. It is also expected to be applicable in systems for searching similar case precedents or constructing event timelines.

3. Methodology

3.1. Overview

The proposed LLM-assisted A2P message forensics methodology comprises three steps as illustrated in Fig. 2: ❶ *data pre-processing*, ❷ *prompt engineering for A2P message classification and named entity recognition*, and ❸ *post-processing for LLM response validation*. For reference, this methodology assumes the availability of a pre-configured environment with a suitable and capable LLM to process input text. Depending on forensic requirements and circumstances, a local or remote LLM engine can be selected.

3.2. Data pre-processing

3.2.1. Email data

Email data is pre-processed to target individual original Multipurpose Internet Mail Extensions (MIME) formatted messages. By extracting only the primary headers (sender, recipient, subject, and date) and body content from individual emails' MIME-formatted messages (which typically have an.EML extension), it is necessary to discard unnecessary data and retain only the core content that is likely to contain forensically useful named entities.

3.2.2. Messenger data

Chat logs from messaging services, particularly incoming messages, undergo pre-processing to separate them into individual messages. This step ensures clarity and structure in the processing workflow, making it easier to analyze each message independently. Specifically, each incoming message is treated as a distinct input to the LLM, allowing for precise and targeted analysis.

3.3. Prompt engineering for A2P message forensics

The next step involves crafting prompts tailored for A2P message forensics. Specifically, two distinct personas are defined (Stöckli et al., 2024): one designed to differentiate between A2P (application-to--person) and P2P (person-to-person) messages, and the other focused on extracting key named entities from classified A2P messages to infer user behavior. The prompts should be carefully engineered to guide the model in recognizing the semi-structured nature of A2P messages, ensuring accurate and meaningful analysis.

3.3.1. A2P message classification

Pre-processed messages are classified as either A2P or P2P using a specially crafted prompt. Figs. 3 and 4 illustrate examples of the prompt and its corresponding response for processing a real-world A2P message. As shown in this example, the prompt consists of the following components:

- Context: Clarification that the model is tasked with identifying and classifying A2P messages
- Examples: Including specific A2P keywords such as *order*, *receipt*, *shipped*, *confirmation*, *departure time*, and more
- Constraints: Explicitly excluding promotional or marketing content not triggered by user actions

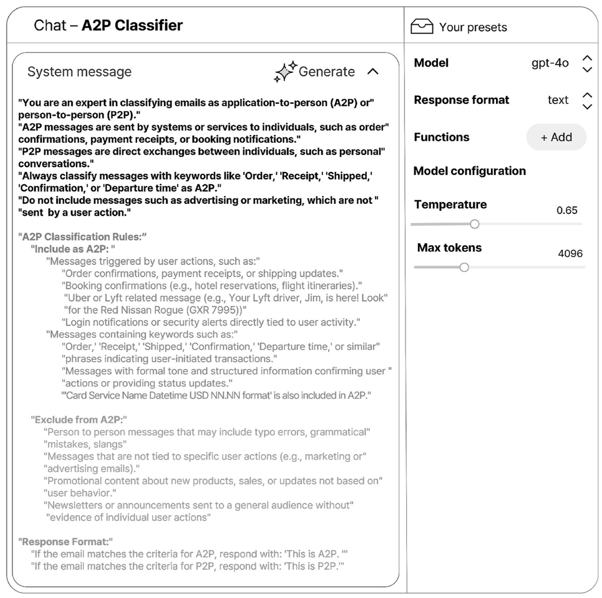


Fig. 3. Prompt specially crafted for A2P message classification.

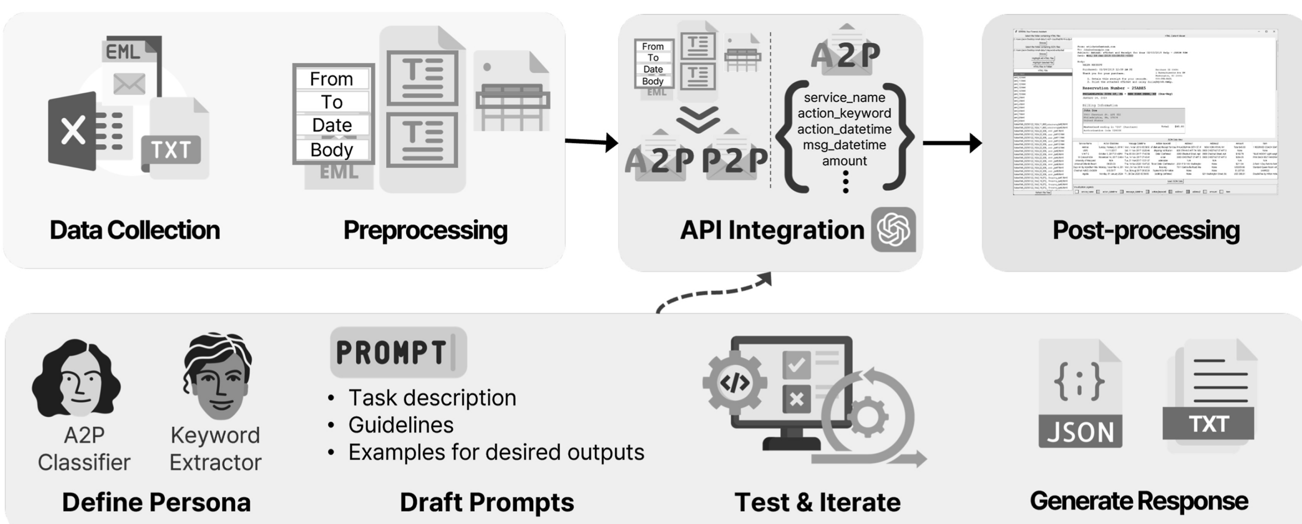


Fig. 2. Overall workflow of our proposed methodology for LLM-assisted A2P message forensics.

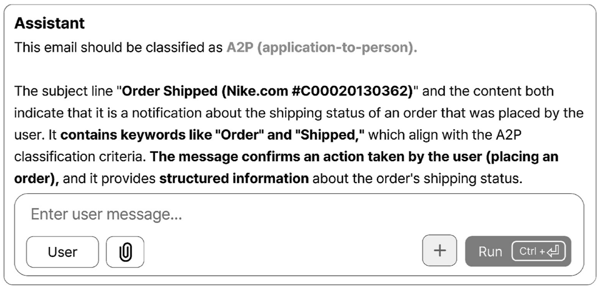


Fig. 4. Example of LLM's response as a result of A2P message classification.

Using this prompt, messages containing transaction history, order confirmations, and payment receipts are identified as A2P. If the applied LLM possesses a high level of contextual understanding, we can expect even long or complex content to be classified relatively accurately. As a point of reference, the following sections will demonstrate that this approach works effectively with one of the state-of-the-art LLMs. Special attention is given to distinguishing between actionable A2P messages-triggered by user behavior, such as orders, payments, reservations, and login attempts-and non-actionable messages, such as marketing emails, product announcements, and other promotional content.

For reference, to ensure accuracy, the prompt provided to the LLM includes detailed guidelines on both actionable A2P messages and excluded types, such as non-actionable promotional content not tied to user behavior. Additionally, the prompt specifies a clear response format: if a message meets the A2P criteria, the LLM responds with "This is A2P." For messages classified as P2P, the response is "This is P2P." This structured approach facilitates precise identification of A2P messages, streamlining the classification process and ensuring consistent outputs.

3.3.2. A2P named entity recognition

After pre-processed messages are classified into either A2P or P2P, the second prompt is used to extract key information from A2P messages. The prompt guides the LLM in accurately retrieving forensically valuable named entities and formatting them in JSON for ease of analysis. Figs. 5 and 6 show examples of the prompt and its corresponding response for processing an actual A2P message. The retrieved data includes the following items:

- Service Name: Identifying the entity providing the service (e.g., Uber, Aribnb, Hotels.com, Apple, etc.)

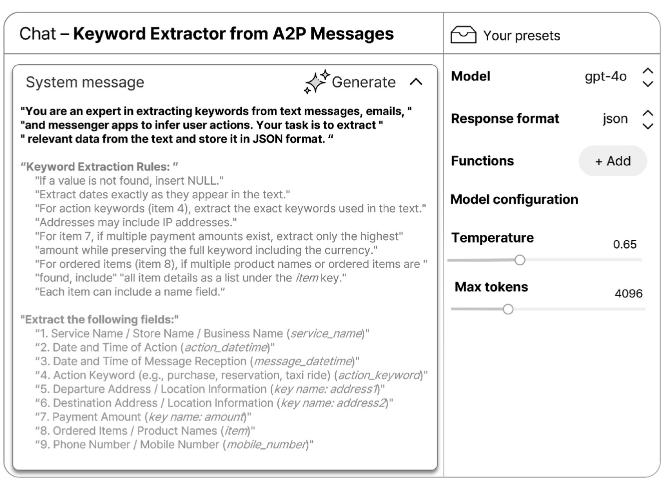


Fig. 5. Prompt specially crafted for A2P-related named entity recognition.

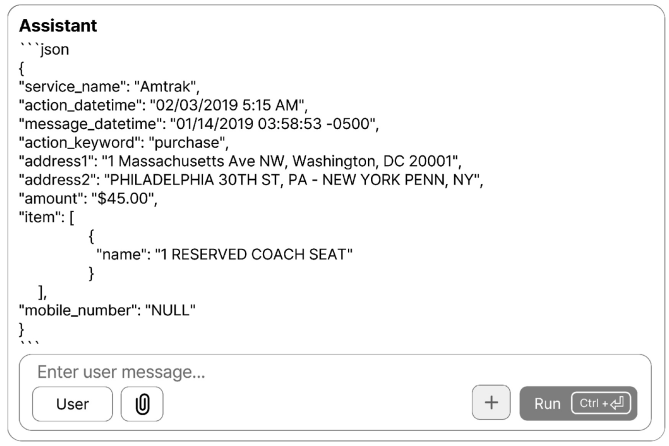


Fig. 6. Example of LLM's response as a result of A2P-related named entity recognition.

- Action Datetime: Timestamp when the user likely performed the action
- Message Datetime: Timestamp indicating when the message was sent to the user
- Action Keyword: Keywords related to the user-triggered action (e.g., booking, confirmation, e-ticket, payment, etc.)
- Address Information: Source and destination addresses, including IP addresses if applicable
- Item Details: Specific products or services mentioned in the message
- Payment Amount: Total cost of the item or service
- Mobile Number: A phone number associated with the service or the user

3.3.3. Considerations for designing prompts

When designing a prompt for classifying messages, the definition of A2P messages should be explicitly stated, emphasizing that they originate from businesses, services, or automated systems in response to user actions. Since the methodology classifies messages strictly as either A2P or P2P, the prompt must also clearly define exclusions to ensure that marketing emails, general announcements, and system-generated notifications unrelated to specific user actions are not misclassified as A2P. Additionally, the response format should be structured—e.g., "This is A2P" or "This is P2P"—to enable automated processing and prevent ambiguous model outputs.

For A2P named entity recognition, the prompt should be carefully designed to guide the model in accurately identifying and structuring relevant information from A2P messages. Handling ambiguous cases is another critical aspect—certain abbreviations (e.g., "AA" for American Airlines) or mnemonic phone numbers (e.g., "800-USA-RAIL") may not represent actual entities. Overall, the model should be guided to differentiate structured data from informal text, ensuring the named entity to extract properly.

3.4. Post-processing for LLM response validation

The LLM processes the previously introduced prompts to generate responses containing forensically valuable named entities extracted from target A2P messages. However, interpreting data through AI models requires caution in digital forensic investigations, as these models may inadvertently distort the facts represented by the data. Therefore, it is essential to establish effective methods for investigators to verify the validity of LLM-generated outputs.

In our methodology, the extracted information is stored in JSON format to facilitate seamless downstream processing, as shown in Fig. 6.

This structured output allows extracted keywords to be directly highlighted within the original message, enhancing the interpretability and reliability of the processed data. By visualizing keywords in context, investigators can validate the accuracy of the extracted information, ensuring a robust and trustworthy analysis.

4. Implementation

In this section, we present the design and development of SERENA (Systematic Extraction and Reconstruction for Enhanced A2P Message Forensics) (Kim, 2025), a proof-of-concept tool that automates the extraction, classification, and analysis of A2P messages using the proposed methodology.

4.1. Design of SERENA

4.1.1. Experimental setup

In this study, we leverage OpenAI's GPT-4o to analyze A2P messages in an automated and efficient manner. GPT-4o was chosen for its enhanced reliability, creativity, and ability to handle nuanced instructions compared to GPT-3.5. Additionally, GPT-4o supports a context window of up to 128,000 tokens, significantly larger than GPT-3.5's 4096-token limit, enabling the processing of longer and more complex messages (TechCrunch, 2023). A Python-based pipeline (Python 3.12.8) was developed to handle incoming text messages and emails, extract key information, and convert the unstructured data into a structured JSON format.

4.1.2. Architecture

SERENA supports various input formats, including emails, Excel files, and messenger chat logs. Its pre-processing module efficiently handles the supported formats. Email data is parsed to extract key headers (e.g., From, To, Subject, Date) and body content, while irrelevant metadata is discarded to reduce processing costs. Chat logs are split into individual messages, each saved as a TEXT file for further analysis. This tool automates the extraction, classification, and highlighting of semi-structured data from A2P messages, providing forensic investigators with a user-friendly interface and robust processing capabilities. The modular design of the pre-processing module ensures scalability, enabling future support for additional data formats.

The architecture of SERENA consists of three main processing modules integrated into a unified pipeline: (1) Data Pre-processing module handles the conversion of raw input data, such as emails, text messages, and chat logs, into a LLM-readable TEXT format for downstream analysis. (2) Message Classification module utilizes OpenAI's GPT-4o to classify messages as either application-to-person (A2P) or person-to-person (P2P). The classification criteria focus on transactional, user-triggered actions while filtering out non-actionable messages like marketing emails. This classification ensures that only relevant messages advance to the keyword extraction module. Finally, (3) Keyword Extraction module extracts forensically valuable named entities from A2P messages, including service names, timestamps, payment amounts, and action keywords. These are saved in a JSON format to enable seamless integration with other processes.

For more details on the keyword extraction module, the current version of SERENA extracts some representative fields of A2P messages including *service_name*, *action_datetime*, *message_datetime*, *action_keyword*, *address1*, *address2*, *amount*, *item* and *mobile_number*. The extracted data is saved as a JSON format as shown in Fig. 7, ensuring traceability and seamless integration with visualization components.

4.2. Execution and outputs

Once the classification and extraction processes are complete,


```
{
  "service_name": "United Airlines",
  "action_datetime": "Sun, Oct 15, 2023",
  "message_datetime": "Fri, 15 Sep 2023 23:58:30 -0400",
  "action_keyword": "purchase",
  "address1": "Quebec City, QC, CA  (YQB)",
  "address2": "Washington, DC, US  (IAD)",
  "amount": "730.32 USD",
  "item": [
    {
      "name": "Flight 1 of 2 UA3610"
    },
    {
      "name": "Flight 2 of 2 UA1050"
    }
  ],
  "mobile_number": null,
  "source_path": "…/A2P-classified-text/a2p_eml_4.txt"
}
```


Fig. 7. Key information in an actual transportation-related A2P message extracted by SERENA with GPT-4o in JSON format.

SERENA generates both raw data outputs and a structured forensic review format. This section describes how SERENA renders its outputs and provides forensic insights.

The tool's output consists of HTML-rendered A2P messages and corresponding JSON files containing extracted data as shown in Fig. 7. Extracted keywords are highlighted within the original HTML-rendered messages using color-coded legends: service name in *yellow*, action datetime in *light green*, message datetime in *light blue*, and action keyword in *pink*, as illustrated in Fig. 8. These highlighted HTML files provide a visual representation of extracted data for forensic review, while a tabular JSON viewer displays all extracted fields, as shown in Fig. 9. This visual format enables forensic analysts to efficiently scan and interpret key information.

5. Evaluation

This section outlines the evaluation framework used to assess the effectiveness of our methodology. We define key performance metrics such as Accuracy, Precision, Recall, and F1-score, which provide a quantitative assessment of classification performance. Below, we describe the calculations and results for our proposed methodology.

5.1. Evaluation metrics

To evaluate the performance of our proposed A2P and P2P classification and keyword extraction methods, we use standard metrics, including accuracy, precision, recall, and F1-Score.

- **Accuracy** measures the proportion of correctly classified instances among all cases:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Precision** (Positive Predictive Value) measures the proportion of correctly identified positive instances among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

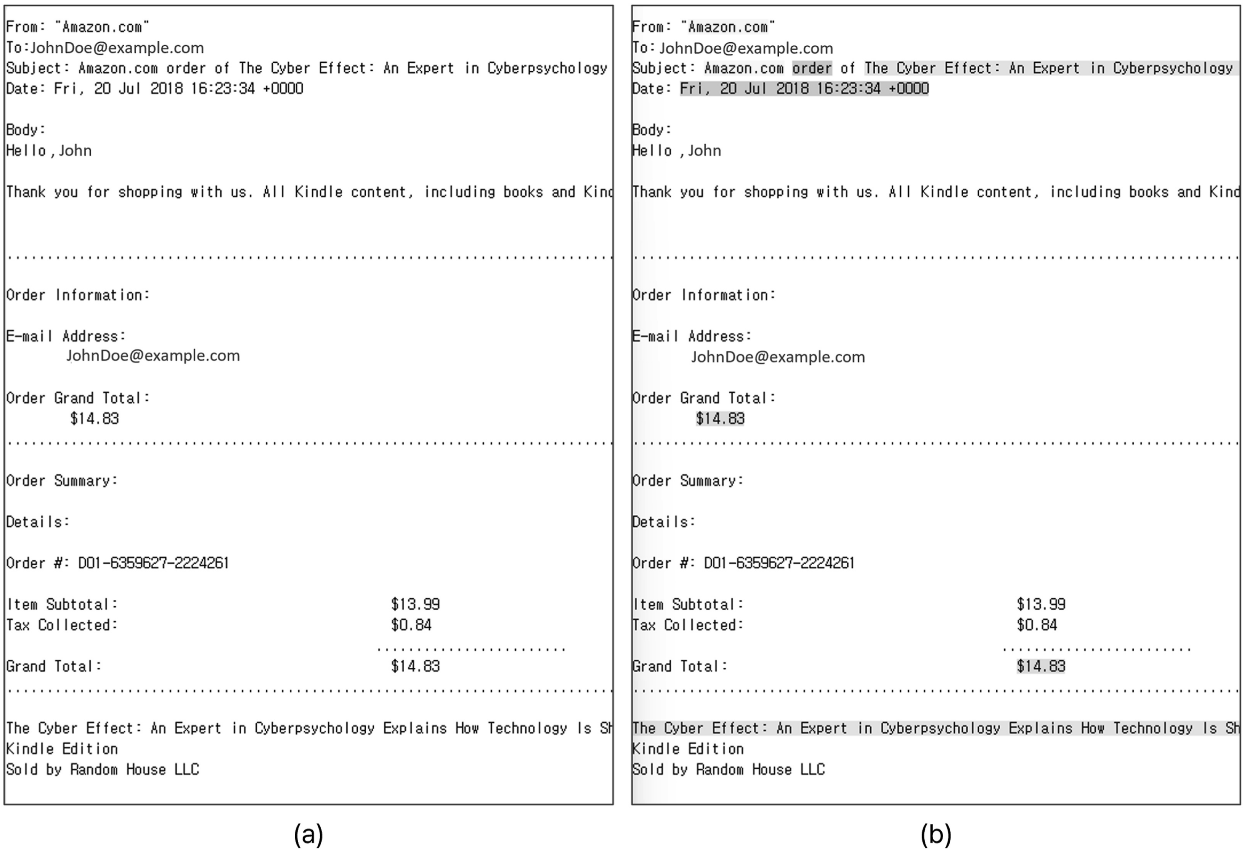


Fig. 8. A screenshot illustrating SERENA's feature to output results that can be highlighted and compared to the original message to assist in determining the validity of the information identified by the LLM.

- **Recall** (Sensitivity) measures the proportion of correctly identified positive instances among all actual positives:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

- **F1-Score** provides a harmonic mean between Precision and Recall:

$$F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

For the metrics of extracting keywords from A2P messages, True Positives (TP) indicate correctly extracted keywords, while False Negatives (FN) represent missed relevant keywords, impacting recall. False Positives (FP) denote incorrectly extracted keywords, which affect precision.

5.2. *Classification of A2P messages*

5.2.1. *Dataset preparation*

We constructed three evaluation datasets—ground truth, augmented, and unseen—based on 50 distinct real-world A2P messages collected from various services across five categories (Order, Transaction, Booking, Notification, and Misc.). These messages, ranging from 11 to 5,620 characters, were manually curated and serve as the basis for the evaluation of the model. Since A2P messages are generated only through actual service usage, the dataset reflects realistic, service-driven communication essential for evaluation. Appendix A provides a detailed overview of the dataset used in this study.

The Ground Truth Dataset. The ground truth dataset comprises 25 A2P and 25 P2P messages that were carefully curated and manually labeled, providing a reliable baseline for performance evaluation.

The Augmented Dataset. The augmented dataset consists of 25 augmented A2P messages and 25 reconstructed P2P messages, both derived from original A2P messages. In the context of A2P message classification, the impact of noisy data—such as mislabeled or ambiguous messages—is relatively limited due to the semi-structured nature of A2P communications. Instead, this dataset simulates edge cases and borderline scenarios to better evaluate the model's robustness against challenging inputs.

Since A2P messages typically have somewhat structured content, we apply augmentation methods that preserve the transactional nature while introducing variations. Each A2P message undergoes one randomly selected augmentation, ensuring realistic diversity without altering the fundamental A2P nature:

- Numerical substitution: Modifying transaction details such as order IDs, tracking numbers, or timestamps
- Synonym replacement: Substituting formal terms with similar expressions (e.g., "receipt" → "invoice")
- Structural variation: Reordering sections within the message body while maintaining logical flow
- Formatting adjustments: Introducing different email signature formats or minor text layout changes
- GPT-based paraphrasing: Utilizing GPT model to paraphrase sections while retaining core meaning

To generate augmented P2P messages, we transform existing A2P messages by restructuring key details such as reservation times, locations, and transaction information into a more conversational and informal format. This process aims at simulating real-world human

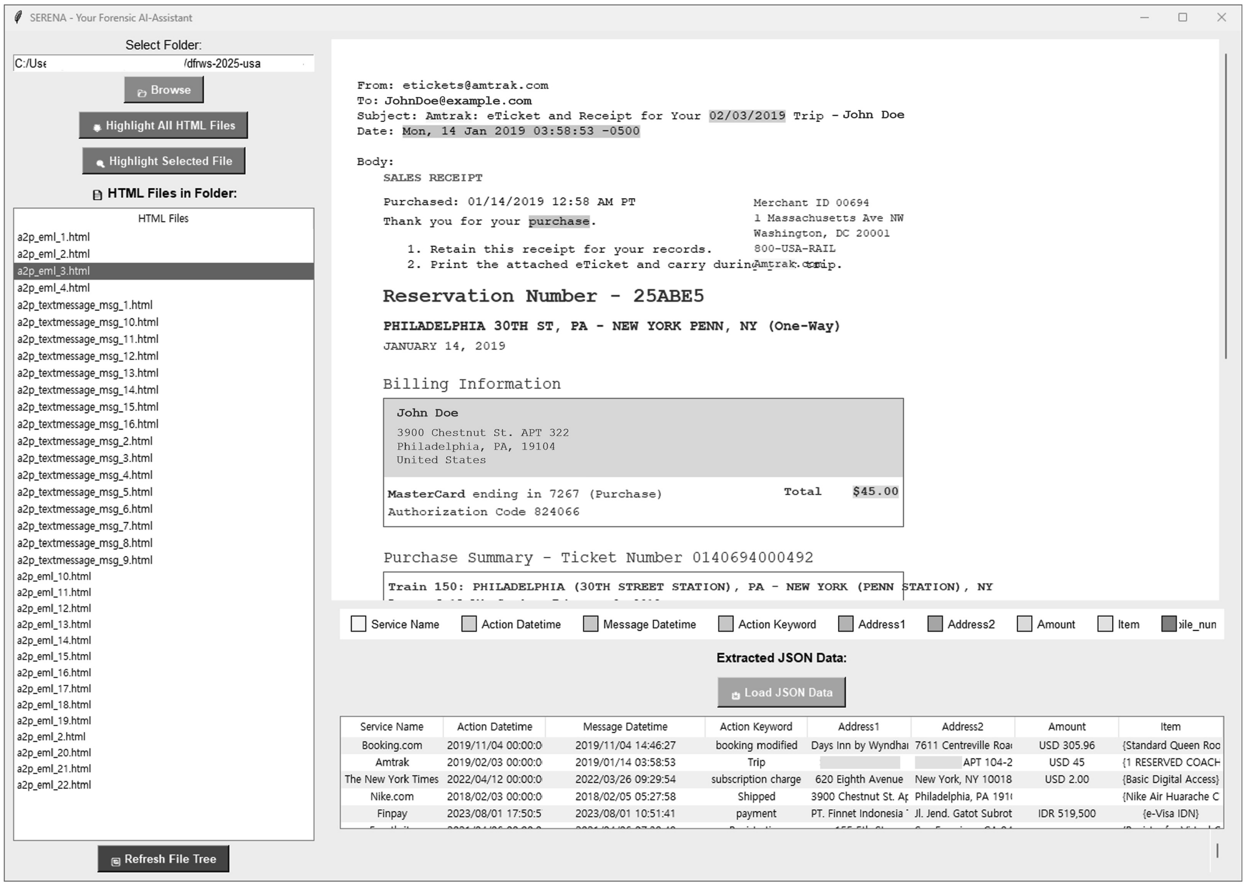


Fig. 9. Main interface of the proof-of-concept tool SERENA and results of processing sample data.

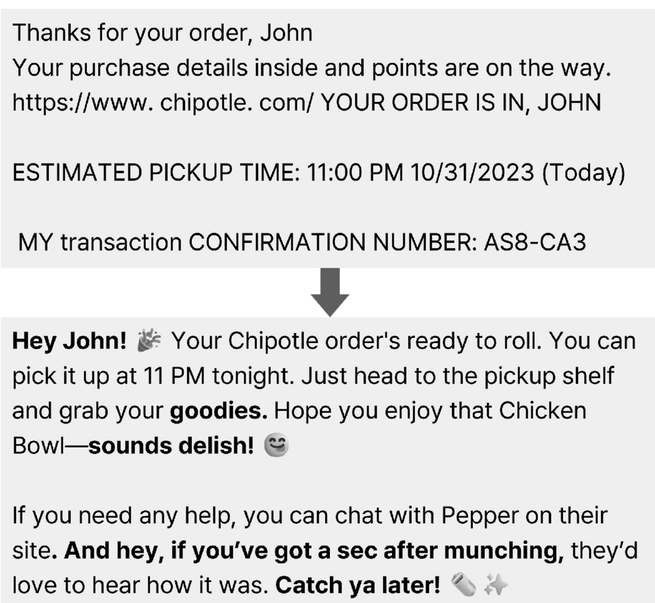


Fig. 10. An augmented P2P Message reconstructed from an original A2P Message.

interactions and ensuring that messages retain the original intent while reflecting the casual tone and structure of P2P communication. Fig. 10 shows an example of the augmented message reconstructed from an A2P message. Techniques included:

- Synonym replacement: Substituting informal words with casual alternatives (e.g., "okay" → "k")
- Sentence reordering: Changing sentence order to simulate natural variation
- Casual element injection: Injecting casual elements like "uh", "hmm", or minor spelling mistakes
- Emoji insertion: Adding emojis to simulate casual digital conversations
- GPT-based paraphrasing: Rewriting messages with slightly different phrasing while keeping the meaning intact

The Unseen Dataset. The unseen dataset consists of 25 A2P and 25 P2P messages that were entirely excluded from the training and validation processes. This evaluation aims to measure the model's ability to generalize to previously unseen, real-world inputs.

5.2.2. Experimental results

The evaluation results highlight the reliability of the proposed method for distinguishing between A2P and P2P messages as listed in Table 2. The classification model demonstrated strong performance across all datasets, achieving 98 % accuracy on the ground truth dataset, with high precision (96.15 %) and recall (96.15 %), ensuring reliable identification of A2P messages. The augmented dataset showed slightly

Table 2
Performance evaluation results of A2P message classification.

Category	Accuracy	Precision	Recall	F-1 Score
Ground Truth	98 %	96.15 %	96.15 %	96.15 %
Augmented Dataset	92 %	86.2 %	100 %	100 %
Unseen Dataset	94 %	92.3 %	96 %	94.1 %

Table 3
Performance evaluation results of A2P named entity recognition across different categories.

Category	Precision	Recall	F1-score
Booking	98 %	98 %	98 %
Miscellaneous	98 %	100 %	99 %
Notification	100 %	100 %	100 %
Order	98 %	100 %	99 %
Transaction	98 %	100 %	98 %


```
{
    "service_name": "Amtrak",
    "action_datetime": "02/03/2019",
    "message_datetime": "Mon, 14 Jan 2019 03:58:53 -
     0500",
    "action_keyword": "eTicket and Receipt",
    "address1": "PHILADELPHIA 30TH ST, PA",
    "address2": "NEW YORK PENN, NY",
    "amount": "$45.00",
    "item":  [
            {"name": "1 RESERVED COACH SEAT"}
      ],
    "mobile_number": "800-USA-RAIL",
    "source_path": "…/Amtrak_eTicket…02_03_2019.txt"
}
```


Fig. 11. An example of incorrect named entity extraction.

lower precision (86.2 %), indicating some false positives, but achieved perfect recall (100 %), meaning all A2P messages were correctly identified. Finally, the unseen dataset maintained a balanced performance with 94 % accuracy, 92.3 % precision, and 96 % recall, implicating the proposed approach to generalize effectively to new data while maintaining strong classification reliability.

5.3. *Keyword extraction from A2P messages*

5.3.1. *Dataset preparation*

To assess the performance of our keyword extraction methodology, we evaluated its effectiveness on five distinct message categories: *Order*, *Transaction*, *Booking*, *Notification*, and *Miscellaneous*. Each category contained 10 messages, totaling 50 A2P messages.

The extracted keywords were compared against manually annotated ground truth data, and the model's precision, recall, and F1-score were computed. The evaluation dataset consists of real-world A2P messages sourced from emails, text messages, and an instant messaging platform. For each message, the model was tasked with extracting up to nine keywords relevant to transaction details, service providers, timestamps, addresses, and other key information.

5.3.2. *Experimental results*

The performance evaluation across message categories shows consistently high accuracy. As listed in Table 3, the model achieves 98 % or higher in precision, recall, and F1-score, demonstrating strong capability in extracting relevant forensic keywords. The *Notification* category achieves a perfect 100 % across all metrics, highlighting the ease of processing structured messages. *Miscellaneous*, *Order*, and *Booking* categories also show strong performance, though minor variations in recall suggest the presence of edge cases. In particular, *Booking* shows slightly lower recall (98 %), possibly due to ambiguous formats such as mnemonic phone numbers. For example, as shown in Fig. 11, the model interpreted "800-USA-RAIL" in a standard phone number format, though it is a mnemonic. These results support the model's robustness in extracting forensic data.

5.4. *Discussion*

In practical forensic investigations, the ability to extract named entities from A2P messages provides investigators with direct access to case-relevant evidence such as transaction amounts, reservation details, timestamps, and service identifiers. For instance, SERENA can identify a timestamped payment to a specific vendor, which may align with a suspect's alibi or contradict a stated timeline. This capability is particularly valuable in digital forensics where user behavior reconstruction plays a critical role in event analysis. By automatically highlighting key elements—such as action times, addresses, and amounts—within the original message context, the system enables investigators to swiftly assess the relevance of each communication, reducing manual review time and increasing evidentiary precision.

6. Conclusions and future directions

In this study, we proposed a methodology for tracking user actions by extracting forensically valuable information from A2P (application-to-person) messages. Our approach was evaluated in two stages: (1) distinguishing A2P messages from non-A2P messages and (2) extracting key information from A2P messages.

By leveraging LLMs with prompt engineering, the method enables accurate extraction of transaction details, locations, and timestamps, helping investigators reconstruct user behavior and verify digital evidence. Although performance slightly declined in augmented and unseen datasets, the results confirm the model's robustness and practical applicability.

This study highlights the forensic potential of A2P messages as a reliable source of timestamped, structured user activity. The use of LLMs enables scalable and adaptable entity extraction without service-specific retraining, which is especially beneficial in diverse and evolving service environments. However, limitations remain in handling domain-specific abbreviations, non-standard formatting, and edge cases such as mnemonic phone numbers. Addressing these challenges is essential to further improve accuracy and ensure consistent performance in complex forensic scenarios.

Our future work aims at enhancing the adaptability of LLMs to domain-specific abbreviations and improving the precision of extracted values. Furthermore, we plan to extend the extracted information using Open-Source Intelligence (OSINT) to enrich forensic investigations. In addition, analyzing email attachments, embedded URLs within messages, and linked external data sources would be helpful to uncover additional contextual insights. By integrating these elements, we expect to provide a more comprehensive understanding of user activities and enhance the forensic value of A2P message analysis.

Acknowledgements

This work was supported by Police-Lab 2.0 Program (www.kipot.or.kr) funded by the Ministry of Science and ICT(MSIT, Korea) & Korean National Police Agency (KNPA, Korea) [Project Name: Research on Data Acquisition and Analysis for Counter Anti-Forensics/Project Number: 210121M07].

Appendix A. A2P Dataset

Table A.4
A detailed overview of the A2P dataset generated and used in our study. It groups messages by service type and lists representative service providers along with example messages for each category. The example messages include placeholders (e.g., *FullName, ActionDate*) that indicate dynamic information that varies per message instance.

Category	Example Services	Example Message
Order	Papa Johns, Chipotle, Amazon, Nike, University, ETS, Weee!, Walmart, Universal Orlando, Mytheresa	**Nike:** Tracking Number *{TrackingNumber}* IT'S ON ITS WAY. Shipping to: *{FullName} {Address} {OrderItem}* Order Date: *{ActionDate}* **Amazon:** Hello *{Full Name}*, Thank you for shopping with us. Order Details: Order *{OrderNumber}*, placed on *{ActionDate}*, Item: *{OrderItem}*, Total: *{TotalAmount}* **Chipotle:** Your order is in, *{Full Name}*. Estimated pickup time: *{ActionDate}*. Order Details: *{Name of Menu}*, Amount: *{TotalAmount}*, Payment Method: *{Apple Pay}*. **University:** Order summary, Name: *{FullName}*, Payment method: *{PaymentMethod}*, Transcripts ordered: *{#}*, Total cost: *{TotalAmount}*, Delivery address: *{Address}*. **Universal Orlando Resort:** Your order was successfully processed. Order Confirmation Thank you for your Universal Orlando Resort™ order. ORDER CONFIRMATION NUMBER: *{ConfirmationNumber}* Item Purchased *{ItemName}* Price *{TotalAmount}* GuestName *{FullName}* Billing Information *{Address}* **Mytheresa:** Order Confirmation *{OrderNumber}* Dear *{FullName}*, Thank you for shopping with us. Please find the summary of your order below.Item(s) ordered *{OrderItem} {TotalAmount}* **Walmart Bakery:** Your custom cake order is ready for pickup. Please pick it up by *{ActionDate}*.
Transaction	Apple, Airbnb, Finpay, Uber, Crypto Exchange, Venmo, GoldCard, Booking.com, Payco, PrestigeCard	**Finpay:** Thank you, we have received your payment. Order Summary: e-VISA IDN. Amount: *{TotalAmount}*, Transaction Date: *{ActionDate}*, Card: *{CardNumber}*. **Airbnb:** Airbnb Receipt. Receipt ID: *{ReceiptID}*, Accommodation Date: *{Accomodation Date}*, Fee Details: *{Total Amount}*. **Apple:** Your receipt from Apple APPLE ID *{UserAccount}* DATE *{ActionDate}* BILLED TO *{Address}* TOTAL *{TotalAmount}* **Venmo:** Venmo You paid *{ReceiverName}* Transfer Date and Amount: *{TransferDate} {TransferAmount}* Completed via a bank transfer from your Pncbank, National Association account ending in *{4 digits Card Number}* **Booking.com:** Reservation Number: *{ReservationNumber}* Payment Completed. Hello, *{LastName}* Your reservation fee for *{Item}* has been successfully processed. *{ActionDate} {TotalAmount}* **GoldCard:** [Web Sender] GoldCard *{CardNumber}* Approval *{TotalAmount} {PaymentDate} {MerchantName}* **Crypto Exchange:** Your Deposit is complete. Hello *{UserAccount}*, Your deposit is complete to your account. Amount: *{Amount}* CompletionDate: *{ActionDate}* **Uber:** Here's your receipt for your ride, *{Name}*, Total: *{TotalAmount}*. You rode with *{DriverName}*, Distance: *{Distance}*, Departure: *{Departure}*, Arrival: *{Arrival}*.
Booking	Dental Clinic, Expedia, Amtrak, Spirit Airlines, Railclick, DoubleTree Hilton, DaysInn, Airbnb	**Dental Clinic:** *{ActionDate}* - Appointment Confirmation for *{FullName}* at *{DentalClinicName}*. Please bring your ID card. **Days Inn by Wyndham Manassas:** Your booking has been successfully modified. This booking was updated on *{ActionDate} {BookingItem}* price: *{TotalAmount}* **Amtrak:** Reservation Number - *{ReservationNumber} {Departure} {Destination}* Billing Information *{Address} {PartialCardNumber} {Total Amount} {ActionDate}* **Expedia:** Your Expedia Booking: *{BookingNumber}*, *{BookingItem}*, Check-in: *{ActionDate}*, Price summary: *{TotalAmount}*. **railclick:** Hi *{FullName}*, your booking is confirmed! *{Departure} {DepartureDate} {Arrival} {ArrivalDate}* **Spirit Airlines:** Your mobile boarding pass is below. Confirmation Number: *{ConfirmationNumber}*, Departing: *{Departure}*, Arrival: *{Arrival}*.
Notification	Magnet Webinar, USPS, Virgin Australia, Eventbrite, NY Times, Government Notices, Airbnb	**MagnetWebinar:** This is a reminder that *{WebinarName}* is scheduled for *{ActionDate}*. **Amazon:** Shipped: Your Amazon package with *{OrderItem}* will be delivered *{ActionDate}* **American Airlines:** Flight AA1796, from CLT to PHL on *{ActionDate}*, departure time has changed. New time of departure is 11:40 PM. Check airport monitors for updates. **Virgin Australia:** Hi,*{FirstName}*,you've successfully checked-in. Thank you for choosing to travel with Virgin Australia. Booking Ref: *{Departure} {Arrival}* Boarding *{BoardingTime}* Departs *{DepartureTime}* **Eventbrite:***{FirstName}*, You've got tickets! Order total: *{TotalAmount}* Order Summary *{ActionDate} {FullName} {OrderItem}* Thank you for registering for *{EventName}* **Amazon:** Your package with *{OrderItem}* will be delivered by *{ActionDate}*.
Miscellaneous	Chope, Google, AuroraRetail, FTK, ExpressVPN, University	**Google:** Someone else used your Google password for account *{UserAccount}*. **Chope:** Thank you for visiting Mari Beach Club. Dear.*{FullName}* Thank you for visiting us at Mari Beach Club, we hope you enjoyed your time and had a great experience with us. **AuroraRetail**: You have logged in from a new device. Hello, *{FullName}*Your account has been logged in from a new device. Recent Login History ID *{UserAccount}* Device *{UserDevice}* Time *{ActionDate}* Location: *{Nation (IP Address)}* **Rainbow University:** Dear *{FullName}*, Congratulations on your admission to the Rainbow University and welcome from the Admission office! The forms linked in this email should be completed by all students. **ExpressVPN:** Welcome to ExpressVPN, *{FirstName}*! Your account details: Email: *{UserAccount}*, Plan: *{OrderItem}*, Price: *{TotalAmount}*.

References

Adkins, J., Bataineh, A.A., Khalaf, M., 2024. Identifying persons of interest in digital forensics using NLP-based AI. Future Internet 16 (11).

Chen, Z., Yang, Y., Chen, L., Wen, L., Wang, J., Yang, G., Guo, M., 2017. Email visualization correlation analysis forensics research. In: 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud), pp. 339–343.

Chernyshev, M., Baig, Z., Doss, R.R.M., 2024. Towards Large Language Model (LLM) Forensics Using LLM-Based Invocation Log Analysis, pp. 89–96.

Cho, S.-H., Kim, D., Kwon, H.-C., Kim, M., 2024. Exploring the potential of large language models for author profiling tasks in digital text forensics. Forensic Sci. Int.: Digit. Invest. 50, 301814.

Egersdoerfer, C., Zhang, D., Dai, D., 2023. Early exploration of using ChatGPT for log-based anomaly detection on parallel file systems logs. In: Proceedings of the 32nd International Symposium on High-Performance Parallel and Distributed Computing. HPDC '23. Association for Computing Machinery, New York, NY, USA, pp. 315–316.

Ghafarian, A., Mady, A., Park, K., 2020. An empirical analysis of email forensics tools. Int. J. Netw. Secur. Appl. 12 (3).

Goel, D., Ahmad, H., Jain, A.K., Goel, N.K., 2024. Machine Learning Driven Smishing Detection Framework for Mobile Security.

Hina, M., Ali, M., Javed, A.R., Srivastava, G., Gadekallu, T.R., Jalil, Z., 2021. Email classification and forensics analysis using machine learning. In: 2021 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing. Scalable Computing & Communications, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/IOP/SCI), pp. 630–635.

K, S.E., Sakshi, Wadhwa, M., 2023. Enhancing digital investigation: leveraging ChatGPT for evidence identification and analysis in digital forensics. In: 2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 733–738.

Kim, J., Park, J., Lee, S., 2023. An improved IoT forensic model to identify interconnectivity between things. Forensic Sci. Int.: Digit. Invest. 44, 301499.

Kim, J., 2025. SERENA. https://github.com/jieonk/SERENA, 2025-04-15.

Magnet Forensics, 2018. Magnet.AI — A Minute with Magnet.

Michelet, G., Breitinger, F., 2024. ChatGPT, Llama, can you write my report? An experiment on assisted digital forensics reports written using (local) large language models. Forensic Sci. Int.: Digit. Invest. 48, 301683.

Nguyen, T.T., Wilson, C., Dalins, J., 2023. Fine-tuning Llama 2 Large Language Models for detecting online sexual predatory chats and abusive texts. arXiv preprint arXiv: 2308.14683.

Oh, D.B., Kim, D., Kim, D., Kim, H.K., 2024. volGPT: evaluation on triaging ransomware process in memory forensics with Large Language Model. Forensic Sci. Int.: Digit. Invest. 49, 301756.

Park, Y., Park, R.S., Kim, H., 2024. Key information extraction for crime investigation by hybrid classification model. Electronics 13 (8).

Pirzada, S., Rahman, N.H.A., Cahyani, N.D.W., Othman, M.F., 2023. A survey of forensic analysis and information visualization approach for instant messaging applications. Int. J. Adv. Comput. Sci. Appl. 14 (2).

Rai, S., Bhattarai, S., Dhar, P., Upadhyaya, B., Sharma, K., Gurung, S., 2023. Analyzing social media texts for suicidal risk identification using Natural Language processing. In: 2023 9th International Conference on Signal Processing and Communication (ICSC), pp. 227–231.

Scanlon, M., Breitinger, F., Hargreaves, C., Hilgert, J.-N., Sheppard, J., 2023. ChatGPT for digital forensic investigation: the good, the bad, and the unknown. Forensic Sci. Int.: Digit. Invest. 46, 301609.

Shahbazi, Z., Byun, Y.-C., 2022. NLP-based digital forensic analysis for online social network based on system security. Int. J. Environ. Res. Publ. Health 19 (12).

Stöckli, L., Joho, L., Lehner, F., Hanne, T., 2024. The personification of ChatGPT (GPT-4)—understanding its personality and adaptability. Information 15 (6).

TechCrunch, 2023. OpenAI releases GPT-4, a multimodal AI that it claims is state-of-the-art. https://techcrunch.com/2023/03/14/openai-releases-gpt-4-ai-that-it-claims-is-state-of-the-art/.

Tejaswini, V., Babu, K.S., Sahoo, B., 2024. Depression detection from social media text analysis using Natural Language processing techniques and hybrid deep learning model. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 23 (1).

Twilio Inc., 2024. A2P (Application-to-Person) SMS messaging. https://www.twilio.com/docs/glossary/what-a2p-sms-application-person-messaging.