USENIX

THE ADVANCED COMPUTING
SYSTEMS ASSOCIATION

Integrating Large Language Models into Security Incident Response

Diana Kramer, *Google;* Lambert Rosique, *DataPhant;* Ajay Narotam,
Elie Bursztein, Patrick Gage Kelley, Kurt Thomas, and Allison Woodruff, *Google*

https://www.usenix.org/conference/soups2025/presentation/kramer

This paper is included in the Proceedings of the
Twenty-First Symposium on Usable Privacy and Security.

August 11–12, 2025 • Seattle, WA, USA

ISBN 978-1-939133-51-9

Open access to the Proceedings of the
Twenty-First Symposium on Usable Privacy and Security
is sponsored by USENIX.

Integrating Large Language Models into Security Incident Response

Diana Kramer
Google

Lambert Rosique
DataPhant

Ajay Narotam
Google

Elie Bursztein
Google

Patrick Gage Kelley
Google

Kurt Thomas
Google

Allison Woodruff
Google

Abstract

Incident response is a manually-intensive process whereby security analysts detect and respond to security events. In this study, we explore whether large language models (LLMs) can fully automate—or otherwise assist with—the final step of an incident response investigation: summarizing findings for stakeholders, auditors, and legal experts. We run a series of experiments with 18 security analysts and 50 real-world incidents to understand (1) whether LLMs can autonomously reason about security events and produce high-quality summaries; (2) whether LLMs can collaboratively assist security analysts with summarization; and (3) what overall benefits and risks security analysts foresee with integrating LLMs into incident summarization. We find that current LLMs may lack the security reasoning necessary to operate autonomously, producing summaries that omit critical details in 35% of cases and/or inject factual inaccuracies in 42% of cases. However, when used collaboratively, LLMs reduce the effort required from analysts to produce a summary, while improving the readability and consistency of summaries. We explore opportunities for improving the security reasoning of LLMs as well as other potential applications for incident response.

1 Introduction

Within enterprise settings, *detection and response*[1] teams are responsible for detecting and responding to security events—such as network intrusions, compromised devices, data breaches, or insider risks. The *security analysts* involved in this process manually investigate user reports and automated alerts (e.g., from endpoint detection systems or network scanners) to identify the validity, scope, and severity of a purported anomalous activity [9, 31]. Through the course of their investigation, security analysts document hypotheses and analysis steps, implement mitigations or remediations, and preserve evidence related to impacted systems or data [11].

With the recent emergence of more powerful AI systems including large language models (LLMs), researchers have begun to explore how to accelerate aspects of incident response. Most studies have focused on automating *detection*: improving the accuracy and prioritization of security alerts to minimize false positives [22, 29], or streamlining log analysis with transformer architectures [3, 12, 20, 23, 37]. Conversely, *response* remains manually intensive, requiring scarce time and expertise from security analysts.

In this paper, we study the experiences and attitudes of security analysts towards LLMs fully automating, or otherwise assisting with, the final stage of response: summarizing an incident to communicate findings with affected stakeholders, auditors, and legal experts. This activity condenses facts and details produced throughout an investigation to succinctly capture (1) the security event that triggered the investigation; (2) actions taken throughout the investigation, including any deployed mitigations or remediations; and (3) details on data or systems accessed or impacted by an attacker. Quality here is crucial: summaries carry business and legal ramifications. They are also used to educate other security analysts on emerging threats and best practices. We consider three research questions tied to this summarization effort:

RQ1: Can LLMs autonomously reason about security events to produce a comprehensive, factually accurate incident summary?

RQ2: Does assistance from an LLM help a security analyst improve the quality or velocity of their incident summaries?

RQ3: What benefits or risks do security analysts foresee when engaging with LLMs as part of incident summarization?

[1] The community also uses the term digital forensics and incident response (DFIR).

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

To conduct our study, we recruited 18 Google security analysts and examined 50 real-world incidents related to cloud intrusions, exposed credentials, malware, phishing, and coin mining attacks on Google's employees and enterprise environments. Across multiple experiments, we asked participants to explain their preference between two side-by-side summaries of the same incident, where a summary might be human-authored, autonomously generated by an LLM, or collaboratively generated by an LLM and human. We find that participants preferred human-only summaries 128 times (61%) over autonomously generated summaries 83 times (39%). This stemmed from LLMs omitting important investigation details in 73 cases (35%) and introducing factual inaccuracies in 89 cases (42%) when acting autonomously. Collaboratively generated summaries were more promising: participants preferred these in in 116 cases (77%), versus human-only summaries in 17 cases (11%).

Our findings suggest that the LLMs used in our experiments may lack the security reasoning necessary to operate fully autonomously in incident response settings, though models are rapidly advancing. Collaboration appears more appropriate in the near-term: LLMs reduce the effort required from security analysts, while also improving the consistency and readability of the summaries. We discuss further opportunities for advancing autonomous summarization in incident response, such as through grounding to improve factuality [15], or relying on fine-tuning [14] or retrieval augmented generation [24] to expand a model's security domain knowledge. Summarization represents just the first step to assisting security analysts with incident response. Improvements in reasoning may unlock more sophisticated tasks, such as recommending next steps in an investigation or helping to interpret anomalous security events, all of which ultimately help analysts more rapidly respond to incidents.

2 Related Work

Applying AI to detection and response. Previous applications of AI towards detection and response have primarily focused on advancing anomaly detection, improving the quality of alerts, and streamlining log analysis. Examples include analyzing document access patterns, SQL queries, and RPC logs to identify insider risk [18]; applying unsupervised learning to middle box logs to assist with ranking alerts [22]; and combining information from multiple detection systems to improve the quality of alerts [29] or to reduce alert fatigue due to false positives [13]. Other research thrusts have focused on how to represent logs more efficiently to enable more complex analysis, relying on fine-tuning or training new transformer models [3, 12, 20, 23, 37]. Ultimately, all of these detection systems feed into the need for response, where summarization plays a final role.

Evaluating the security and privacy knowledge of LLMs. As a preliminary step to understanding the deployability of AI in security and privacy applications, researchers have benchmarked whether state-of-the-art LLMs can correctly respond to closed-form or short answer security and privacy questions. These include questions sourced from standards, certifications, and research papers [2, 26, 33], challenges from capture the flag competitions [34], crowdsourcing [17], or questions derived from industry knowledge bases [8]. For some of these benchmarks, state-of-the-art LLMs achieve over 90% accuracy on 10,000 questions [33]. Our study explores a more complex task than knowledge alone: whether an LLM can sift through a large context window of incident details and yield a coherent summary of the events that unfolded.

Evaluating the quality of AI-generated summaries. Researchers have leveraged LLMs to summarize information from a variety of sources including news articles [36], source code [1, 38], and legal documents [10]. Evaluating the quality of these summaries falls into one of two modes: human judgement or automated scores. Common criteria for human judgement include *coherence* (e.g., well-structured), *consistency* (e.g., factually accurate), *fluency* (e.g., free from formatting or grammatical errors), and *relevance* (e.g., captures most important details) [19]. We utilize similar criteria for our evaluations, except scoped to the security and privacy domain. Automated techniques instead rely on having a "golden reference" that generated summaries are compared against using metrics such as ROUGE [25], METEOR [4], BLEURT [32], BERTScore [35], and more. While we have access to existing human summaries in our study, part of our research explores how LLMs can improve upon these summaries. As such, we rely on qualitative side-by-side comparisons from human domain experts to provide more nuance compared to automated metrics.

3 Methodology

In collaboration with Google's detection and response team, we ran a series of experiments with 18 security analysts working at Google to study whether LLMs can automate, or otherwise assist with, summarizing incidents. We briefly overview the investigation details that feed into these summaries, how we recruited participants, and the overall design of our study to evaluate the quality of human and LLM summaries as well as the attitudes of security analysts towards LLMs playing a role in incident response.

3.1 Incident Response Workflow

We outline the flow of an incident response investigation in Figure 1. While our study focuses on participants from Google, the practices they follow are similar to those published by NIST [9].

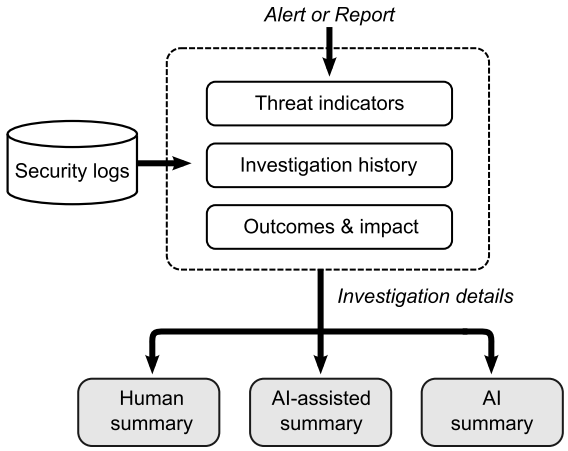

Figure 1: The incident response workflow is capstoned by summarizing the threat indicators that triggered an investigation, the investigative steps taken, and any relevant outcomes or impacts of the incident. The summaries highlighted in grey form the basis of our experiments.

An investigation—or *ticket*—is typically triggered by an automated alert (e.g., from endpoint, network, or log monitoring systems) or an employee-initiated report of a suspicious event (e.g., receiving a phishing email). Throughout the course of an investigation, a security analyst (or multiple analysts)[2] will record investigation details in the ticket related to (1) threat indicators (e.g., device identifiers, IP addresses, URLs); (2) a history of actions taken (e.g., hand-offs, status updates, remediation steps taken); and (3) details related to the outcome and impact of the incident (e.g., disabling an account or device, whether an attacker gained access to sensitive data), or whether the trigger was a false positive. Given the real-time nature of investigations, security analysts will document hypotheses and supporting evidence as they work, leading to the possibility of contradictory or outdated statements.

At the conclusion of an investigation, all of these details are collated into a final summary by the security analyst. This summary represents a mundane, but critical, step. Summaries are used to inform impacted stakeholders, to educate other detection and response team members, and for auditing and verification. Inaccurate or inconsistent summaries thus pose both a business and legal risk.

The purpose of our study is to explore whether LLMs can entirely automate, or otherwise assist with, the generation of summaries—saving a security analyst time and/or effort, while simultaneously improving quality and/or consistency. We focus on summarization versus more complex incident response tasks—such as recommending investigation steps or interpreting threat indicators—as all of the context necessary for a summary exists within the investigation details of a ticket associated with an incident, and for the near-term practical impact of automated or assisted summarization.

[2]Typically, less complex investigations will involve a single security analyst, while more complex investigations will involve many.

3.2 Recruiting & Participants

We recruited 18 security analysts from Google's detection and response team, conscientious of the time it took for these 18 security analysts to participate in our study. We engaged all participants from July 2024–December 2024 based on their availability. All participants received $350, or equivalent local value, for their time regardless of how far they progressed in our study. Participation was entirely voluntary. Of our participants, 2 had between 1–5 years of experience with security incidents, 4 between 6–10 years of experience, and 12 had 10+ years of experience.

3.3 Sourcing Incident Summaries

Human summaries. Our team selected 50 incidents with human-authored final summaries for use in the study. These 50 events were drawn from incidents that security analysts had previously investigated as part of their day-to-day work. All of these incidents were "frozen" in the sense that the ticket associated with each investigation was closed and no new additional details or changes to the summary were pending. To ensure a diversity of summaries, the team used random, stratified sampling across five categories of security events,[3] selecting ten summaries per category from all the tickets from 2022–2023 as detailed in Table 1. This earlier time period reduces any risk of LLMs being involved in incident investigations and thus contaminating our study; it also minimizes the risk of recency bias in the event a participant previously investigated the exact incident. We refer to this dataset as *human summaries*.

AI summaries. Using the same 50 incidents, we provided all of the investigation details associated with each incident (excluding the final summary) to Gemini 1.5 Flash to summarize.[4] We disabled any logging to prevent sensitive incident details being saved outside our experiment environment. We used a prompt that Google's detection and response team had previously experimentally iterated on to produce a summary (see Appendix A). The prompt emphasizes using language accessible to a junior security analyst, legal team member, or non-security stakeholder. The prompt evaluates semi-structured details associated with an incident that are programmatically represented in an XML format, as shown in

[3]While in practice there are other categories of security events, these reflect the most common events, covering roughly half of the investigations conducted by the security analysts participating in our study.

[4]Neither Gemini 2.0, 2.5, or related Deep Research models were available at the time of our investigation. Other potential models from companies such as OpenAI or Anthropic were not feasible, given the sensitivity of the investigation details associated with each incident. Other models would likely perform similarly per prior benchmarks of security knowledge [7].

Incident	Security event trigger	Paraphrased incident summary
Cloud intrusion	Alert that a Google Cloud instance has been compromised (e.g., due to insecure credentials or a software vulnerability).	An attacker brute forced a weak SSH credential to access [Virtual Host Identifier]. The attacker infected the host with malware that then engaged in port scanning. [Virtual Host Identifier] was disabled and the owner deleted the instance.
Coin mining	Alert that a Google Cloud instance or device was mining cryptocurrency, typically due to a malware infection.	[Virtual Host Identifier] was detected due to excessive CPU resource usage and contacting [IP Address]. Device confirmed to be mining cryptocurrency. [Virtual Host Identifier] has been disabled and deleted.
Credential exposure	Alert or report of an employee exposing their credential in clear text.	An employee inadvertently committed a configuration file containing a Google Cloud service account key for [Project Identifier] to GitHub. The configuration file was reverted and the exposed key was disabled. Analysis of the activity logs for [Project Identifier] show no evidence of misuse.
Malware	Alert that an employee device has been infected with malware (e.g., via endpoint detection signals or other detection capabilities).	[Alerting Tool] triggered when two malicious driver files, [Filename] and [Filename], executed on a Windows machine [Device Identifier]. The files hashes matched [specific rule details]. Initial analysis suggested the malware achieved persistence by loading during system startup. The investigation determined the presence and subsequent removal of malicious files.
Phishing	Employee-initiated report that they are the target of a suspected phishing attempt or broader campaign.	Phishing impersonation of [Service Name] using [Impersonated URL]. [User] reports having been successfully phished but no corporate credentials or system were compromised. Reviewing the raw content of the email, the message was sent via noreply@[Authentic URL] using one of [Authentic URL]'s mail servers, indicating their mail server may be compromised. [User] password reset. Only one other employee received the email, but did not respond to it.

Table 1: Five types of incidents included in our study. 10 summaries of each type were selected from real-world incidents occurring from 2022–2023, totaling 50 summaries. We provide a paraphrased incident summary from each type to illustrate the expected level of detail.

Figure 2. These details include a description of what triggered the incident, metadata about the type of incident and timing information, tags applied by the security analyst reviewing the incident for suspected root causes (e.g., misconfigured system, weak password), software details of affected systems, and a detailed log of comments from the security analyst recorded throughout their investigation. We refer to the summaries we produced using LLMs with these details as *AI summaries*.

3.4 Study Procedures

Part I: Comparing human vs. AI summaries. We first asked participants to read the details associated with an incident—accessed via the same tools used in their day-to-day work—and then assess which of two text descriptions, either A or B, best summarized the incident, based on criteria consistent with their training and workplace best practices. One summary was the original human summary written at the time of the incident as described earlier, while the other was entirely AI-generated. Participants also provided an open-ended rationale for their selection (e.g., "Both summaries are not perfect. But B contains statements that are factually wrong like 'misconfiguration was fixed'"). See Appendix B for our full instrument.

In total, every participant was shown 14–15 incidents randomly selected from our pool of 50 possible incidents. We randomized the order of whether a human or AI summary appeared as summary A or B per comparison. The median participant spent 3:23 minutes reviewing each incident and summary, and 56 minutes overall. A total of 15 participants engaged in this experiment, producing 211 preferences and justifications.[5]

Part II: Generating AI-assisted summaries. In the next experiment, we showed participants either a blank space or an initial AI summary and asked them to modify the text to produce a complete, factually correct, and concise summary of an incident as they would in their day-to-day work. When starting from an initial AI summary, we asked participants to describe the changes they made to the starting summary and rate how difficult it was to start from a blank space versus an initial AI summary (with five options ranging from *"Much easier to edit, than starting from scratch"* to *"Much harder to edit, than starting from scratch"*). Based on failure modes that we identified by coding the open-ended responses in Part I (described shortly), we also asked participants the extent to which they needed to add additional details or correct factual errors (with five options ranging from *"A great deal"* to *"Not*

[5] A single participant was able to review only 5 incidents; 4 participants reviewed 14 incidents; and the remaining 10 participants reviewed 15 incidents.

<Title>Abuse verdict for [Google Cloud Instance].**</Title>**
<Metadata>Ticket was filled and submitted on [Timestamp]. It was marked with the label "Cloud Intrusion".**</Metadata>**
<Description>[Detection System] detected abusive behavior on [Virtual Host Identifier].**</Description>**
<Incident Causes>MISCONFIGURATION and WEAK PASSWORD.**</Incident Causes>**
<Software Involved>[System Details]**</Software Involved>**

<Mitigation History>
<Comment>There was a network traffic spike [URL evidence] around [Timestamp]. Running [Mitigation Tool] now.**</Comment>**
<Comment>Confirmed instance is compromised. Disabling the instance and notifying owner.**</Comment>**
<Comment>Machine was accessed via SSH using the root user and a weak password: [URL Evidence].**</Comment>**
<Comment>Malicious cron job was created using [Code Snippet]. It downloaded a file from [IP Address].**</Comment>**
<Comment>Malware engaged in port scanning attack before instance disabled.**</Comment>**
<Comment>Instance deleted by owner. Summary added. Closing ticket.**</Comment>**
</Mitigation History>

Figure 2: Paraphrased details related to an incident, passed to the LLM for summarization.

at all"), as independent questions. See Appendix C for full details.

We provided every participant with 10 randomly selected incidents that they did not see in Part I. Of these, 5 had blank spaces, and 5 contained initial AI summaries. The former served as a control to measure the time difference between starting from scratch, versus with a draft created by an LLM. We refer to to the output of the latter as *AI-assisted summaries* which have human oversight. Writing a summary (including participants familiarizing themselves with an incident's details) took a median of 3:07 minutes per incident, and 39 minutes overall.[6] A total of 15 participants engaged in this experiment, producing 142 summaries.[7]

Part III: Attitudes towards AI-assistance. Immediately after a participant completed Part II, we asked them to respond to a survey that gauged how mentally demanding editing AI summaries was overall, whether they felt it led to higher-quality summaries, and whether they preferred collaborating with AI rather than working on their own, and why. A total of 13 participants completed the survey.[8] See Appendix D for full details.

Part IV: Comparing human vs. AI-assisted summaries. As a final validation, we asked two security analysts who did not participate in the previous experiments to review all incidents and evaluate which of two text descriptions—the original human summary written at the time of the incident and the AI-assisted summary from Part II for the same incident—best summarized the incident, or if they were equally good. The purpose of this evaluation was to gauge if AI-assistance led to higher-quality, equivalent, or lower-quality summaries, based on 75 samples of AI-assisted summaries. We did not engage the earlier participant pool for this experiment to minimize the overhead on participants, and to avoid introducing any bias due to a participant being familiar with incidents from earlier experiments or due to authoring a summary.

3.5 Analysis Approach

We relied on a mixture of quantitative and qualitative approaches to analyze the data produced by each experiment. For closed-form questions from Parts I–IV, we report counts and percentages of ordinal values.

We enrich this based on the open-ended feedback provided by participants in Parts I–III. Here, we used thematic analysis [5] to identify common issues reported by participants pertaining to the quality of a human or AI summary. While there are parallels between the terminology used by participants and the terms of art used by the AI community (see Section 2), we opted to retain the language used by security analysts in both our experiment instructions and in our reported results. We relied on an inductive approach to iteratively develop and refine a codebook [30]. We began by familiarizing ourselves with the data from Parts I–III. Three authors then collaboratively reviewed all open-ended responses to Part I to generate an initial codebook. These same authors then applied this codebook to 15% of the open-ended responses to Part I as a research team, making minor adjustments to the codebook and reaching consensus on all items. Then, two of these authors independently applied the codebook to all responses to Part I, later reconciling all disagreement through discussion.[9] See Appendix E for our codebook. One author subsequently applied a simplified version of the codebook

[6] We suspect the speed here stems from participants not having to compare two summaries, as in Part I, and due to many years of experience producing summaries.

[7] One participant completed 5 summaries, another 7 summaries, and the remaining 13 participants each completed 10 summaries. One participant was new to Part II and did not participate in Part I. This totals 16 participants across Parts I and II, with the final two participants in only Part IV.

[8] Of these, 10 participants completed the survey immediately after the experiment. Another 2 completed it within 10 days, and 1 within a month.

[9] As both coders coded all items and agreed on all the final codes, we do not report inter-rater reliability (IRR).

to Part II, confirming that previously identified codes fully covered the open-ended responses from Part II. Open-ended data from Part III informed our analysis and aligns with our codebook but was not coded due to the sparsity of the data.

Throughout the analysis, the authors coding the data had access to the human and AI summaries seen by participants, or to the summaries generated by participants, and they periodically referred to these summaries to resolve ambiguities in open-ended responses, but these authors did not have access to the underlying tickets seen by the participants due to the sensitivity of the data. We report counts from Part I produced by this coding, along with representative quotes from Parts I-III, throughout our results.

For statistical significance tests, we relied on a binomial regression $Y_i \sim B(n_i, \pi_i)$, using a binary variable to represent a preference for an AI or AI-assisted summary (e.g., 1) or a human summary (e.g., 0)—or the prevalence of a theme (e.g., 1) or absence of a theme (e.g., 0)—and treating the type of incident (e.g., coin mining, malware) as a categorical variable. We selected one category to act as a baseline (or reference) during modeling by sorting the values under measurement and choosing the category associated with one pole of the range. For comparing numeric distributions (e.g., length of summaries, time to complete a summary), we used an independent t-test. We used Python's `statsmodels` for these calculations.

3.6 Ethics

Our study plan was reviewed by experts at Google in domains including ethics, human subjects research, policy, legal, security, privacy, and safety. We note that Google does not require IRB approval, though we adhere to similarly strict standards. The researchers involved in this study never had access to raw incident details. Researchers did have access to summaries; all summaries were produced by Google's detection and response team according to the methods previously outlined. In order to further protect participants and any sensitive business details, we refer to participants only by ID, and omit unique details from quotes as they relate to Google-specific system names or tools.

3.7 Limitations

Our study design carries a number of limitations. The incidents in our dataset do not cover the entire range of possible security, privacy, or safety issues an enterprise may encounter, having been selected from just five categories. As such, our results may not generalize to other categories of incidents or detection and response teams. The incidents we selected may not generalize to incidents seen at other companies. The prompt for summarization was selected by Google's detection and response team, as was the model they interfaced with; other prompts, tool chains, fine-tuning, or more recent LLMs might yield different results. Our study design required participants to familiarize themselves with an incident after-the-fact. In practice, security analysts writing summaries would be intimately familiar with the incident they investigated. This might yield discrepancies compared to an in situ experiment. Likewise, participants in our study may already interact with LLMs as part of their day-to-day work, potentially biasing their opinions on the value it brings to incident response. Finally, the statistical power of the sample sizes of our study are limited; we complement our quantitative findings with qualitative findings. Where feasible, we include measures of statistical significance.

4 Results

Our analysis reveals that the LLMs evaluated in our study fall short of matching the quality of summaries produced by security analysts: in head-to-head comparisons, human summaries were preferred to AI summaries 61% of the time. This stemmed from LLMs omitting critical details, introducing factual inaccuracies, or making other errors in summaries. However, we find that AI *assisting* security analysts yields multiple positive improvements related to both quality and efficiency.

4.1 Evaluating AI Summaries

Based on the results of Part I of our experiments, participants preferred human summaries 128 times (61%), versus AI summaries 83 times (39%). This was dependent on the type of security event involved as shown in Table 2. Malware incidents proved the most difficult for AI to accurately summarize, with participants preferring human summaries in 33 cases (79%, $p = 0.005$). Conversely, participants were evenly split between preferring human or AI summaries for phishing and coin mining incidents. Our coding revealed that participants cited four main factors for their preference for a summary: completeness, factuality, concision, and readability.[10] While such issues are common with AI-generated content [16, 19], we explore their unique impact on incident response. For each factor, we summarize when AI had problematic issues compared to human summaries in Table 2 and when human summaries had issues compared to AI summaries in Table 3.

Completeness. Participants mentioned that AI summaries were incomplete or missing critical details 73 times (35%, Table 2)—particularly for malware incidents ($p < 0.001$). AI summaries could lack relevant technical details used to gauge the severity of an incident:

> *"[Human summary] provides more details around the compromised device including the contents of*

[10] As multiple codes could be assigned, a given incident might have multiple issues, for example, a single incident might have both a completeness issue and a factuality issue.

Incident	Prefer human summary	Prefer AI summary	p	Completeness issue	Factuality issue	Concision issue	Readability issue
Cloud intrusion	29 (67%)	14 (33%)	0.082	10 (23%)	22 (51%)	8 (19%)	0 (0%)
Coin mining	21 (49%)	22 (51%)	–	11 (26%)	18 (42%)	13 (30%)	0 (0%)
Credential exposure	25 (60%)	17 (40%)	0.324	16 (38%)	18 (43%)	10 (24%)	2 (5%)
Malware	33 (79%)	9 (21%)	0.005	26 (62%)	17 (40%)	1 (2%)	1 (2%)
Phishing	20 (49%)	21 (51%)	0.996	10 (24%)	14 (34%)	3 (7%)	2 (5%)
Total	128 (61%)	83 (39%)	–	73 (35%)	89 (42%)	35 (17%)	5 (2%)

Table 2: Summary of whether security analysts preferred human or AI summaries, with 211 comparisons overall. We also break down four issues which were common reasons that security analysts preferred human summaries over AI summaries. P-values (p) indicate whether the preference for human vs. AI summaries is statistically significant across incident types, with coin mining as a baseline.

Improvement over human summary	N	%
Completeness	87	41%
Factuality	9	4%
Concision	23	11%
Readability	29	14%

Table 3: Summary of common reasons that security analysts preferred AI summaries over human summaries.

> *the instance for reference. This information is useful as it has an impact on the potential severity of the incident."* – P1

> *"[Human summary] notes there was no evidence of the key being used, which is relevant and not included in [AI summary]."* – P5

Other issues stemmed from omitting key details about remediation steps or all the relevant outcomes of an investigation:

> *"[Human summary] is better because it provides more details on the remediation steps such as contacting [impacted stakeholders] about updating the machines in [region.]"* – P11

> *"[AI summary] is missing information about the fact that the workaround is published by the software vendor, and is lacking details about the other actions ([submitting] bug to [another team] and broken flows)"* – P9

Conversely, despite these frequent concerns, participants positively indicated that AI summaries *did* have relevant details 87 times (41%, Table 3). In many cases, these details were missing from human summaries:

> *"[Human summary] misses important information such as the attack vector was RDP and that the attacker created multiple accounts [...] [AI summary] also has remediation information which is lacking in [Human summary]."* – P14

> *"[AI summary] is better because it provides more details about the nature of the alert and the remediation steps such as shutting down the infected VM and notifying the VM owner."* – P11

Factuality. AI summaries included imprecise or entirely contrived (e.g., "hallucinated") details 89 times (42%, Table 2). These factual errors occurred for every type of incident, but were most common for cloud intrusions, and least common for phishing, though these differences are not statistically significant (all $p > 0.05$). Some of these errors related to technical subtleties in how incidents unfolded, or cases where steps taken throughout an investigation were misattributed:

> *"[AI summary] makes it sound like the malware was blocked before it ever had a chance to run which is not the case (first stage ran, second stage denied)."* – P1

> *"[AI summary] incorrectly states that the malware is called [identifier]. This is an incident codename, not name of the malware."* – P14

> *"[AI summary] has factually wrong information: 'The [support engineer] immediately advised the user to change their credentials.' This is not stated anywhere in the [incident details] and actually was the recommendation of the [security analyst] working on the case to the [support engineer]."* – P4

More severe factual errors related to invalid investigation conclusions, whereby the AI summary inflated the severity of an incident, or contrived remediation steps:

> *"[AI summary] states a guess at the vector for the key leak."* – P9

> *"[AI summary] is stating that the team is working with the manufacturer, which is not true."* – P6

> *"[The] claim that the 'project owner is working to improve security measures' is not mentioned in the [incident details] at all and is a false statement."* – P14

> *"[AI summary] implies that the machine itself was running an ML model making it seem this machine was more critical then it is."* – P15

Human summaries were not immune from including errors, although this was much more rare. Participants critiqued the factual accuracy of human summaries 9 times (4%, Table 3):

> *"[Human summary] incorrectly states that a Kubernetes cluster was compromised. Only one VM instance in the affected project was compromised, not the whole cluster."* – P11

> *"[Human summary] has what appears to be many hallucinations, such as [tool] being malware, or that [team] did not collect anything.*[11]*"* – P12

Concision. Participants flagged AI summaries as being overly-verbose 33 times (16%) and overly-short 2 times (1%, Table 2). Coin mining ($p = 0.007$), credential exposure ($p = 0.018$), and cloud intrusion ($p = 0.039$) had the most concision issues. Here, participants most commonly cited unnecessary "fluff" generated by the AI:

> *"I think the first sentence of [AI summary] is unnecessary fluff ('which was detected by our security systems'). The fact that an [investigation] was created in general indicates that the activity was detected and this doesn't need to be stated."* – P1

> *"There's zero additional details in B but it's 5x longer for no reason."* – P8

Conversely, participants mentioned that AI summaries provided a more appropriate level of detail compared to human summaries 23 times (11%) (Table 3):

> *"[Human summary] doesn't seem complete/draft and reads more like notes taken from the ticket details. [AI summary] is a concise summary of the ticket and captures the issue, actions taken."* – P13

Readability. Participants cited readability issues with AI summaries just 5 times (2%, Table 2), wherein a summary was difficult to parse or understand. All differences between types of incidents are statistically insignificant ($p > 0.05$). Instead, participants shared that AI summaries were more readable than human summaries 29 times (14%, Table 3). However, this improvement in readability was often outweighed by issues related to completeness or factuality:

> *"[AI summary] has a more readable summary but [Human summary] contains more relevant information [...] [Human summary] is also more nuanced in stating the cause."* – P14

[11] Here, the participant appears to believe that the human summary was actually created by AI, and attributes 'hallucinations' to it.

~~A Google Cloud project was targeted by an intrusion attempt, which was detected by our security systems.~~ The investigation revealed that the ~~project~~ VM was compromised and it was running Wannacrypt malware. The VM was most likely compromised via a weak password for an admin account exposed via [Remote Desktop Protocol]~~,~~ ~~leading to the shutdown and deletion of the affected instance.~~ The ~~incident is now resolved and the~~ project owner ~~has been~~ was notified and the affected VM was deleted.

Figure 3: Example edit applied to an initial AI summary.

Additional details required to be added	N	%
A great deal	7	9%
Quite a bit	13	17%
Somewhat	12	16%
A little	22	29%
Not at all	17	23%
No response	4	5%

Table 4: Extent to which participants reported adding additional details to 75 AI summaries. Values do not add to 100% due to rounding.

4.2 Evaluating AI-Assisted Summaries

Using the data from Part II of our experiments, we analyzed both common types of edits applied by participants to initial AI summaries, as well as measures of the effort involved to bring an AI summary inline with a participant's expectations of a high-quality summary. We show an illustrative example of one edit from a participant in Figure 3.

Editing for completeness. Participants self-reported having to add additional detail *"Somewhat"*, *"Quite a bit"*, or *"A great deal"* for 32 of 75 summaries (43%, Table 4). This is slightly larger, though inline with our earlier results (Table 2). When an initial AI summary was deemed "too generic" or of

Factual errors requiring correction	N	%
A great deal	10	13%
Quite a bit	10	13%
Somewhat	13	17%
A little	11	15%
Not at all	27	36%
No response	4	5%

Table 5: Extent to which participants reported correcting factual errors in 75 AI summaries. Values do not add to 100% due to rounding.

Effort versus starting from scratch	N	%
Much easier to edit	19	25%
Slightly easier to edit	17	23%
Not easier or harder	21	28%
Slightly harder to edit	8	11%
Much harder to edit	5	7%
No response	5	7%

Table 6: Effort spent by participants editing 75 AI summaries. Values do not add to 100% due to rounding.

little to no assistance, participants would simply start from scratch:

> *"Completely replaced the summary as I quickly found it was missing key information"* – P14

> *"I wiped it and wrote my own. It was WAY too verbose saying things like 'intrusion detected by Google Security Systems.' Yeah [obviously]."* – P8

Editing for factuality. Participants self-reported correcting factual errors *"Somewhat"*, *"Quite a bit"*, or *"A great deal"* in 33 of 75 summaries (44%, Table 5). This is inline with our earlier results (Table 2). In some cases, these factual edits were limited:

> *"Added [...] and reworded the mitigation actions as they did not align to the actual steps taken in the ticket."* – P1

> *"Removed an incorrect statement about investigation being ongoing"* – P13

While in others, they were substantial:

> *"The original summary was inaccurate in that: 1. It misidentified affected users as employees when the report stated they were [contractors]. 2. It mistakenly stated one of the affected user was identified. Neither of the two users involved was successfully identified. 3. It mistakenly stated that the ticket was closed because no action was needed when in fact the closing note indicated that investigation was closed due to the lack of supporting data."* – P11

Size of edits. Given an initial AI summary and the final summary produced by a participant after correcting for completeness, factuality, and other issues, we can measure the fraction of the initial AI summary that changed in terms of Levenshtein distance (Figure 4). We find participants changed a median of 85 characters, or 27% of the initial summary. Just

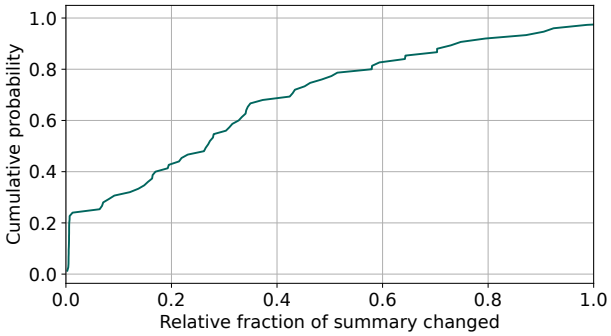

Figure 4: CDF of the relative changes introduced to a summary by participants, measured in terms of Levenshtein distance versus the original length of the summary. Half of summaries required editing 27% of the content.

1 in 4 initial AI summaries required less than a 1% change. These edits had little impact on the overall length of a summary as shown in Figure 5. AI summaries before any edits were a median of 349 characters versus 345 characters after edits ($p > 0.05$). However, these summaries are much longer compared to the original human summaries that had a median of 264 characters ($p = 0.002$). This suggests that participants largely left intact the structure and many incident details introduced through AI assistance.

Effort of edits. On a per summary basis, participants self-reported that it was *"Much easier to edit than starting from scratch"* or *"Slightly easier to edit than starting from scratch"* for 36 of 75 summaries (48%, Table 6). Participants said it was much easier or slightly easier to start from scratch for just 13 summaries (17%). Error margins for a 95% confidence interval are roughly 11%, indicating participants tended to believe AI assistance reduced the effort required.

Overall duration. We measured the relative time it took a participant to edit a summary versus writing a summary from scratch. Participants took on average 250 seconds to edit an initial AI summary, versus 310 seconds from scratch, though these results are not statistically significant for our sample size ($p = 0.208$). As participants had to familiarize themselves

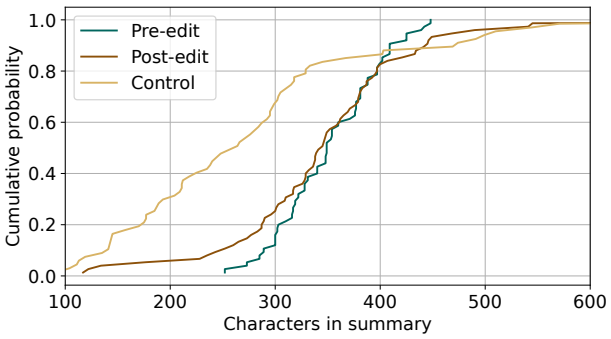

Figure 5: CDF of the overall length of AI summaries before (pre-edit) and after a participant iterated on the summary (post-edit). While the median length of summaries stayed approximately the same, participants often trimmed unnecessary details, while adding more relevant details. AI assistance led to much longer summaries compared to summaries authored by humans alone (control).

with the details of an incident (which would not be true in situ), this familiarization process may represent the bulk of the time spent and one possible reason that AI assistance does not appear to reduce the overall time required to produce a summary.[12]

4.3 Attitudes Towards AI

Our analysis from Part III of our experiments captures the overall impressions of participants towards interacting with AI when constructing summaries. We collate the closed-form responses to three survey questions in Figure 6. Participants were split on whether they felt working with AI yielded higher-quality summaries, with 5 saying quality was higher with AI and 5 saying quality was better from scratch; 3 reported no difference. Here, some participants recognized the benefits of AI to consistency, though cautioned the trade-off was not clearly worth it:

> *"While I feel the structure of summaries consistently can be improved by AI, the potential repercussions from a hallucination adding an inaccurate fact is worrisome."* – P16

Seven participants felt that starting from scratch was more mentally demanding than collaborating with AI, versus just 4 participants who felt interacting with AI was more demanding—a weak indicator that AI streamlined summarization. A consistent theme among unfavorable participants was the added tax of fact checking AI summaries. Some participants presumed this would take less work for incidents they were actively working on—versus our experiment where they had to familiarize themselves:

> *"It's mentally taxing determining whether or not the AI made any factual errors, but if I had all the context before I might be able to pick those up faster/use less of my brain to find them. "* – P15

> *"When the AI gets all the facts right, [AI summarization] works like a charm. On the other hand, when the AI makes things up, it takes more time to edit the incident summary than if I were to start from scratch due to the additional overhead of having to fact check the AI-generated summary"* – P11

Taken as a whole, 7 participants preferred to work with initial AI summaries, while 5 preferred to work alone. Improving the experience for participants would require addressing issues surrounding factuality and completeness:

> *"If I can reasonably trust the quality the AI-generated drafts would be helpful."* – P14

Absent such improvements, one participant felt the benefit of AI was all-or-nothing rather than collaborative:

> *"If [a summary] reads fairly accurate to what I would write I would use it, if it doesn't I would just write from scratch."* – P13

Even were a summary to be free of issues, automation also risked removing a critical time for reflection:

> *"I find the process [of writing summaries] incredibly valuable as it forces me to zoom out of the details of a case and view it at a high level. This helps me verify that I've answered all the original questions from the start of the case, that I followed all investigative threads to a reasonable degree, and everything adds up/makes sense from end to end. It also helps me remember my 'elevator pitch' for each incident I work in just in case I'm asked about it weeks or months later."* – P3

4.4 Final Evaluation

While participants in Part III were roughly split on whether they personally felt AI-assistance actually yielded better quality summaries, we find our final side-by-side evaluation from Part IV provides an alternate picture on the potential benefits of AI-assistance as shown in Table 7. Based on 150 comparisons conducted by two security analysts between an original human summary and the AI-assisted summaries, we find that in 116 cases (77%), the AI-assisted summary was preferred and in another 17 cases (11%) the AI-assisted summary was deemed equivalent in quality to the original summary. In just

[12] We were unable to differentiate time spent writing a summary and time spent examining incident details due to limited instrumentation options with the day-to-day tools used by security analysts.

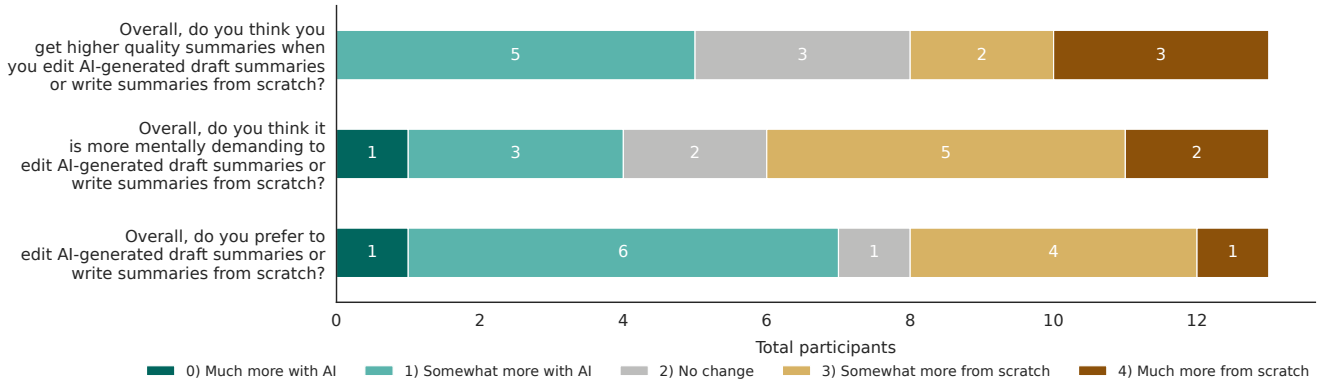

Figure 6: Attitudes of participants towards interacting with AI during summarization. Participants were roughly split on whether AI assistance led to higher-quality summaries, but more preferred working with AI than not.

Incident	Prefer AI-assisted	About the same	Prefer human
Cloud intrusion	30 (75%)	6 (15%)	4 (10%)
Coin mining	17 (85%)	0 (0%)	3 (15%)
Credentials	18 (75%)	4 (17%)	2 (8%)
Malware	27 (79%)	3 (9%)	4 (12%)
Phishing	24 (75%)	4 (12%)	4 (12%)
Total	116 (77%)	17 (11%)	17 (11%)

Table 7: Summary of whether security analysts preferred AI-assisted summaries or human summaries, with 150 comparisons overall. Differences between incident types are not statistically significant (all $p > 0.05$).

17 cases (11%), the original summary was still preferred. This held across incident types, with no statistically significant differences. While we caution there are potential confounding variables—security analysts may be more pressed for time during incidents compared to our experiment; familiarity with incident details may differ from our setup; and biases exist in judging the quality of summaries—our findings suggest that AI-assistance may yield higher-quality summaries.

5 Discussion

Promising results for quality of AI-assisted summaries. AI-assisted summaries were highly rated as compared with original human summaries; in 77% of cases experts preferred the AI-assisted summary and in another 11% of cases the AI-assisted summary was deemed equivalent to the human summary. This suggests that integrating AI assistance in existing incident summarization practices is highly feasible. Fully automated AI summaries performed less well, with participants preferring human summaries 61% of the time versus AI summaries 39% of the time. However, our results should be considered in the context of the usage of Gemini 1.5 Flash and the associated prompt, which leaves ample opportunity for further gains in light of quickly evolving models, prompting strategies, and other technical improvements. In some cases, organizational tolerance for errors may also increase the feasibility of full automation even with higher error rates. Our research outlines an experimental approach for organizations to determine how to strategically use LLMs to assist security analysts with incident summarization.

Bifurcation of attitudes towards AI assistance hinges on factuality. Participants were roughly split on whether they preferred our implementation of AI assistance in their workflows, versus authoring summaries from scratch—with seven favorable participants, and five unfavorable. Based on the feedback provided by participants, this hesitation largely stemmed from the mental effort required to root out factual inaccuracies from an LLM's output, with 59% of summaries requiring at least one factual edit. Participants expressed less concern around addressing issues of completeness. This provides a clear direction for where future improvements to AI-assisted incident summarization will have the largest impact.

Additionally, there is an apparent gap in quality assessment between participants who wrote AI-assisted summaries (with participants roughly split regarding whether they think AI-assisted summaries versus human summaries are of higher quality overall) versus our experiment in Part IV which shows that when independently rated, AI-assisted summaries are considered higher quality 77% of the time. This gap may be due to a variety of causes, for example, those authoring summaries may be highly concerned about potential factual errors that in practice materialize less often or are less severe than they anticipate; this bears further investigation.

Balancing human verification with benefits of AI. Our data illustrates there are clear benefits to integrating AI into

incident response workflows: AI assistance yielded better-structured, more comprehensive summaries in 77% of cases. Yet even if security analysts serve as a guard against factual inaccuracies introduced by LLMs, there is a risk that fatigue may result in some hallucinations—or other errors—slipping through, resulting in a business or legal ramification. This risk is not all together new: junior (or even senior) security analysts may also introduce mistakes—as suspected by participants in 4% of human summaries. One design pattern to mitigate this risk is the use of quality control. Here, one or more security analysts would review a sample of summaries produced via AI assistance (and without) in an ongoing fashion to ensure the quality remains within tolerated bounds set by an organization. More robust controls to mitigate the risks of AI integration remain an open question.

Improving AI summaries for incident response. Best practices for interfacing with LLMs are constantly evolving, leading to a variety of possible improvements to how incidents are summarized. One option would be to fine-tune a state-of-the-art model with information from thousands of incidents, guiding an LLM to the expected structure of incident summaries. Another approach would be to decompose summarization into a series of chain-of-thought tasks where an LLM reasons about the trigger of an incident and the most important actions taken throughout an investigation, before merging the output of each of these reasoning tasks into a final summary. Chain-of-thought reasoning and more sophisticated grounding techniques [15] can also be used for an LLM to self-critique its summary, ensuring that all findings are tied to a statement recorded by a security analyst in an incident's details. The security reasoning of LLMs can also be extended via retrieval augmented generation, similar to Sec-Gemini's approach of extending Gemini's knowledge to include threat intelligence, best practices, and research [7]. Ultimately, more sophisticated models, tool chains, and prompts may obviate the need for such bespoke designs. Our research serves as a resource for model developers to help close these gaps.

AI in the workplace. In our experiments, we largely focused on the creation and quality of incident summaries, and above we discuss ways that LLM technology (model selection, prompt design, etc.) can be improved to further increase the quality of AI-assisted or fully automated AI summaries. More generally, we can consider incident summarization in the broader context of the full set of tasks that security analysts perform in their jobs. Economists, for example, have been exploring how AI and LLMs will (or will not) replace specific tasks or entire job roles and the accompanying effects on productivity, opportunity, and well-being [6, 27, 28]. Partially or fully automating incident summarization may have broader impacts on worker experience and job satisfaction. For example, LLM summarization could allow security analysts more time to perform more fulfilling, expert, or core tasks (e.g., investigating incidents) or alternatively the industry may deem it important for analysts to continue to perform the critical thinking component [21] of summarization. We leave to future work a broader economic, labor-informed assessment of potential impacts of LLM summarization on security analysts' jobs, and best practices for its practical deployment.

6 Conclusion

In this paper, we explored the effectiveness of integrating LLMs into the final stage of incident response: summarizing key findings for stakeholders, auditors, and legal experts. We considered two possible scenarios for how LLMs might interface with a security analyst summarizing an incident: fully autonomously, or in an assistive role. We experimented with each scenario using summaries created with Gemini 1.5 Flash and a bespoke summarization prompt, gathering feedback from 18 security analysts. We found that the LLM in our experiments had limited utility when operating autonomously, with participants preferring human summaries 61% of the time over LLM-generated summaries due to the latter omitting critical details and inadvertently introducing factual inaccuracies. Conversely, participants assessed that collaborative summaries (produced by an LLM and security analyst) were higher quality 77% of the time compared to summaries produced by a security analyst alone. Beyond these quality improvements, 48% of the time participants felt that assistance from an LLM was easier on a per-incident basis than working alone (versus 17% of the time when it was counterproductive). Our findings highlight how LLMs can be integrated into incident response today *with oversight*; while outlining key improvements necessary to achieve full automation.

Acknowledgments

We thank the security analysts who participated in this work for generously sharing their expertise.

References

[1] T. Ahmed, K. S. Pai, P. Devanbu, and E. Barr. Automatic semantic augmentation of language model prompts (for code summarization). In *Proceedings of the IEEE/ACM International Conference on Software Engineering*, pages 1–13, 2024.

[2] M. T. Alam, D. Bhusal, L. Nguyen, and N. Rastogi. CTIBench: A benchmark for evaluating LLMs in cyber threat intelligence. *arXiv preprint arXiv:2406.07599*, 2024.

[3] C. Almodovar, F. Sabrina, S. Karimi, and S. Azad. Logfit: Log anomaly detection using fine-tuned language models. *IEEE Transactions on Network and Service Management*, 2024.

[4] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.

[5] V. Braun and V. Clarke. Conceptual and design thinking for thematic analysis. *Qualitative Psychology*, 9(1):3–26, 2022.

[6] E. Brynjolfsson, D. Li, and L. Raymond. Generative AI at work. Technical report, National Bureau of Economic Research, 2023.

[7] E. Burzstein and M. Tishchenko. Google announces Sec-Gemini v1, a new experimental cybersecurity model. https://security.googleblog.com/2025/04/google-launches-sec-gemini-v1-new.html, 2025.

[8] Center for Internet Security. CIS benchmarks list. https://www.cisecurity.org/cis-benchmarks, 2025.

[9] P. Cichonski, T. Millar, T. Grance, and K. Scarfone. Computer security incident handling guide. https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-61r2.pdf, 2012.

[10] A. Deroy, K. Ghosh, and S. Ghosh. Applicability of large language models and generative models for legal case judgement summarization. *Artificial Intelligence and Law*, 2024.

[11] Google. Data incident response process. https://cloud.google.com/docs/security/incident-response#data_incident_response_process_2, 2024.

[12] H. Guo, S. Yuan, and X. Wu. LogBERT: Log anomaly detection via BERT. In *Proceedings of the International Joint Conference on Neural Networks*, 2021.

[13] W. U. Hassan, S. Guo, D. Li, Z. Chen, K. Jee, Z. Li, and A. Bates. NoDoze: Combatting threat alert fatigue with automated provenance triage. In *Proceedings of the Network and Distributed Systems Security Symposium*, 2019.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] A. Jacovi, A. Wang, C. Alberti, C. Tao, J. Lipovetz, K. Olszewska, L. Haas, M. Liu, N. Keating, A. Bloniarz, et al. The FACTS grounding leaderboard: Benchmarking LLMs' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*, 2025.

[16] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

[17] P. Jing, M. Tang, X. Shi, X. Zheng, S. Nie, S. Wu, Y. Yang, and X. Luo. SecBench: A comprehensive multi-dimensional benchmarking dataset for LLMs in cybersecurity. *arXiv preprint arXiv:2412.20787*, 2024.

[18] A. Kantchelian, C. Neo, R. Stevens, H. Kim, Z. Fu, S. Momeni, B. Huber, E. Bursztein, Y. Pavlidis, S. Buthpitiya, M. Cochran, and M. Poletto. Facade: High-precision insider threat detection using deep contextual anomaly detection. *arXiv preprint arXiv:2412.06700*, 2024.

[19] W. Kryściński, N. S. Keskar, B. McCann, C. Xiong, and R. Socher. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*, 2019.

[20] V.-H. Le and H. Zhang. Log parsing with prompt-based few-shot learning. In *Proceedings of the IEEE/ACM International Conference on Software Engineering*, 2023.

[21] H.-P. H. Lee, A. Sarkar, L. Tankelevitch, I. Drosos, S. Rintel, R. Banks, and N. Wilson. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems*, 2025.

[22] J. Lee, F. Tang, P. M. Thet, D. Yeoh, M. Rybczynski, and D. M. Divakaran. Sierra: Ranking anomalous activities in enterprise networks. In *Proceedings of the IEEE European Symposium on Security and Privacy*, 2022.

[23] Y. Lee, J. Kim, and P. Kang. LAnoBERT: System log anomaly detection based on BERT masked language model. *Applied Soft Computing*, 2023.

[24] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 2020.

[25] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[26] Z. Liu. SecQA: A concise question-answering dataset for evaluating large language models in computer security. *arXiv preprint arXiv:2312.15838*, 2023.

[27] A. McAfee, D. Rock, and E. Brynjolfsson. How to capitalize on generative AI. *Harvard Business Review*, 2023.

[28] National Academies of Sciences, Engineering, and Medicine. Artificial intelligence and the future of work. https://doi.org/10.17226/27644, 2024.

[29] T.-M. Roelofs, E. Barbaro, S. Pekarskikh, K. Orzechowska, M. Kwapień, J. Tyrlik, D. Smadu, M. Van Eeten, and Y. Zhauniarovich. Finding harmony in the noise: Blending security alerts for attack detection. In *Proceedings of the ACM/SIGAPP Symposium on Applied Computing*, 2024.

[30] J. Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE Publications Ltd, 2009.

[31] D. Schlette, M. Caselli, and G. Pernul. A comparative study on cyber threat intelligence: The security incident response perspective. *IEEE Communications Surveys & Tutorials*, 2021.

[32] T. Sellam, D. Das, and A. P. Parikh. BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

[33] N. Tihanyi, M. A. Ferrag, R. Jain, T. Bisztray, and M. Debbah. CyberMetric: A benchmark dataset based on retrieval-augmented generation for evaluating LLMs in cybersecurity knowledge. *arXiv preprint arXiv:2402.07688*, 2024.

[34] A. K. Zhang, N. Perry, R. Dulepet, J. Ji, J. W. Lin, E. Jones, C. Menders, G. Hussein, S. Liu, D. Jasper, P. Peetathawatchai, A. Glenn, V. Sivashankar, D. Zamoshchin, L. Glikbarg, D. Askaryar, M. Yang, T. Zhang, R. Alluri, N. Tran, R. Sangpisit, P. Yiorkadjis, K. Osele, G. Raghupathi, D. Boneh, D. E. Ho, and P. Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. *arXiv preprint arXiv:2408.08926*, 2024.

[35] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*, 2019.

[36] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 2024.

[37] X. Zhang, Y. Xu, Q. Lin, B. Qiao, H. Zhang, Y. Dang, C. Xie, X. Yang, Q. Cheng, Z. Li, J. Chen, X. He, R. Yao, J.-G. Lou, M. Chintalapati, F. Shen, and D. Zhang. Robust log-based anomaly detection on unstable log data. In *Proceedings of the European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2019.

[38] Y. Zhu and M. Pan. Automatic code summarization: A systematic literature review. *arXiv preprint arXiv:1909.04352*, 2019.

Appendix

See below for the prompt provided to the LLM to summarize an incident and instructions and mock instrument provided to participants throughout our different experiments.

A Summary prompt

You are an Incident Security Engineer working at Google for 10 years.

Task:
Summarize the following Security Incident for your junior colleague in two sentences. Avoid technical words, code, dates, sensitive data in your summary and make sure it answers the questions "what happened", "what was discovered during the investigation", "what was done or is being done to resolve it" and "what is the current status or next steps if it is still ongoing".

Do not split your answer in multiple sections but instead write one short summary merging all the information from those questions.

Examples:

- "A Googler reported that they were targeted by a phishing attempt via LinkedIn. The phishing attempt was made by a person claiming to be from a legitimate charity and asked the Googler to donate money. The Googler reported the profile to LinkedIn and blocked the contact. No further action was required."
- "A coin mining alert was triggered for project [Project Identifier]. Investigation determined that the original compromise vector was a Jupyter notebook that was exposed to the internet with no authentication. The instance was stopped and deleted. There was no sensitive data in the project."

Incident you have to summarize in two to three sentences:
<Incident>[Incident details]</Incident>

Remember that a good summary tells in concise human understandable terms what happened, what is the impact on the data, what we did to investigate and mitigate it and the applied solutions all in just two or three sentences.

Stick to facts, do not make assumptions or recommendations and do not specify dates like exactly when something happened.

B Part I: Experiment design

B.1 Instructions

1. Read through the [incident] ticket.
2. Read the 2 summaries.
3. Evaluate which summary is better, based on your expertise as a security engineer.

For example, a good summary is:

- Complete
- Factually correct
- Concise
- Free of jargon

A good summary also specifies:

- The security issue that triggered the incident
- The action(s) taken to resolve the issue
- The impact of the issue
- The current status or any next steps

4. Explain why you believe one summary is better than the other.

B.2 Instrument

[Ticket link | Summary A | Summary B]

Which summary is better, A or B? [SELECT ONE]

- A
- B

Why do you think that summary is better? Please be specific. [OPEN ENDED]

C Part II: Experiment design

C.1 Instructions

1. Read through the [incident] ticket.
2. Write a new summary, or edit the provided summary.

A good summary is:

- Complete
- Factually correct
- Concise
- Free of jargon

A good summary also specifies:

- The security issue that triggered the incident
- The action(s) taken to resolve the issue
- The impact of the issue
- The current status or any next steps

3. Fill out [additional detail columns], if applicable.

C.2 Instrument

Write or edit your summary here:

[Ticket link | Editable Summary]

Describe the changes you made to the original summary. [OPEN ENDED]

To what extent did you need to add additional details to the original summary? [SELECT ONE]

- A great deal
- Quite a bit
- Somewhat
- A little
- Not at all

To what extent did you need to correct factual errors in the original summary? [SELECT ONE]

- A great deal
- Quite a bit
- Somewhat
- A little
- Not at all

How much easier or harder was it to edit the summary, compared to starting from scratch? [SELECT ONE]

- Much easier to edit, versus starting from scratch
- Slightly easier to edit, versus starting from scratch
- Not easier or harder
- Slightly harder to edit, versus starting from scratch
- Much harder to edit, versus starting from scratch

D Part III: Experiment design

D.1 Instructions

Please answer the following questions about your experience.

D.2 Instrument

Overall, do you think it is more mentally demanding to edit AI-generated draft summaries or write summaries from scratch? [SELECT ONE]

- Much more mentally demanding to edit AI-generated draft summaries
- Somewhat more mentally demanding to edit AI-generated draft summaries
- No substantial difference
- Somewhat more mentally demanding to write summaries from scratch
- Much more mentally demanding to write summaries from scratch

Overall, do you think you get higher quality summaries when you edit AI-generated draft summaries or write summaries from scratch? [SELECT ONE]

- Much higher quality when I edit AI-generated draft summaries
- Somewhat higher quality when I edit AI-generated draft summaries
- No substantial difference
- Somewhat higher quality when I write summaries from scratch
- Much higher quality when I write summaries from scratch

Overall, do you prefer to edit AI-generated draft summaries or write summaries from scratch? [SELECT ONE]

- Strongly prefer to edit AI-generated draft summaries
- Somewhat prefer to edit AI-generated draft summaries
- No preference
- Somewhat prefer to write summaries from scratch
- Strongly prefer to write summaries from scratch

In your daily work, would you prefer to edit AI-generated draft summaries, or would you prefer to write summaries from scratch? Why? [OPEN ENDED]

Please share any other comments you have (optional). [OPEN ENDED]

E Codebook

Completeness. A participant makes a comment about a summary having necessary detail(s), or otherwise is missing important detail(s) core to the interpretation of an incident.

- AI summary has relevant information
- AI summary is missing relevant information

Factuality. A participant makes a comment about the details of a summary being correct, or otherwise containing passages that are explicitly false or unverifiable.

- AI summary has factually accurate information
- AI summary has factually inaccurate information
- AI summary has unverifiable claims not present in the original ticket

Concision. A participant makes a comment about the succinctness or verbosity of a summary, or having an appropriate or inappropriate level of detail.

- AI summary has unnecessary or extraneous information
- AI summary is short, and this is viewed as a positive
- AI summary is short, and this is viewed as a negative
- AI summary is long, and this is viewed as a positive
- AI summary is long, and this is viewed as a negative

Readability. A participant makes a comment about the readability of a summary, such as it being cohesive or well-written; or being incoherent.

- AI summary is easy to read
- AI summary is difficult to read

AI Summary Quality. A participant makes a comment about the overall quality of an AI summary, with the quality being:

- Excellent
- Good
- Ok
- Not good
- Bad

Human Summary Quality. A participant makes a comment about the overall quality of a human summary, with the quality being:

- Excellent
- Good
- Ok
- Not good
- Bad