Evaluating LLMs for Police Decision-Making: A Framework Based on Police Action Scenarios

Sangyub Lee[1, 2], **Heedou Kim**[1, 3], **Hyeoncheol Kim**[1*]
[1]Department of Computer Science and Engineering, Korea University, Seoul, Republic of Korea
[2]Korea National Police University, Asan, Republic of Korea
[3]Police Science Institute, Asan, Republic of Korea
{yubii2, heedou123, harrykim}@korea.ac.kr

Abstract

The use of Large Language Models (LLMs) in police operations is growing, yet an evaluation framework tailored to police operations remains absent. While LLM's responses may not always be legally "incorrect", their unverified use still can lead to severe issues such as unlawful arrests and improper evidence collection. To address this, we propose PAS (Police Action Scenarios), a systematic framework covering the entire evaluation process. Applying this framework, we constructed a novel QA dataset from over 8,000 official documents and established key metrics validated through statistical analysis with police expert judgements. Experimental results show that commercial LLMs struggle with our new police-related tasks, particularly in providing fact-based recommendations. This study highlights the necessity of an expandable evaluation framework to ensure reliable AI-driven police operations. We release our data and prompt template.

Data & Code —
https://github.com/Heedou/PASFramework

Introduction

Recently, police officers face significant operational challenges. Given the high workload caused by frequent overtime, crime scene exposure, and a substantial number of cases, officers often experience significant stress, which increases the risk of errors in police operations or delays in case processing (Stotland 1991; Tan et al. 2022; Vila 2006). To address these challenges, the integration of Large Language Models (LLMs) as auxiliary tools has become increasingly common, presenting the potential to reduce time and human resource consumption. For example, police officers can now leverage state-of-the-art LLMs for tasks such as traffic accident analysis, automated police report generation, automatic phishing detection, and criminal investigations (Kim et al. 2024; Sarzaeim, Mahmoud, and Azim 2024; Tong et al. 2024; Halford and Webster 2024; Adams 2024; Jamal and Wimmer 2023; Kim and Lim 2022) This integration fosters expectations that, in the long run, LLMs will contribute to a safer public security service (Kim et al. 2024; Sarzaeim, Mahmoud, and Azim 2024).

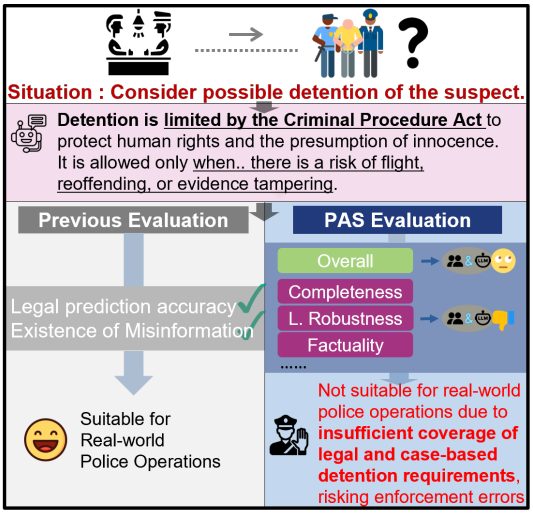

Figure 1: LLMs are increasingly used to support police. However, existing evaluation focus only on information accuracy (Kim et al. 2024; Fei et al. 2023; Hwang et al. 2022), which risks indiscriminate use in real-world scenarios by overlooking key police-specific considerations. Our framework offers guidance for improving LLM from a policing perspective.

Yet, rigorous validation is essential for applying LLMs in specialized domains. Recent studies have conducted both quantitative and qualitative evaluations of LLM performance across various fields, including law and healthcare (Liu et al. 2024a,b; Fei et al. 2023; Hwang et al. 2022). While these models sometimes achieve high accuracy, practical limitations remain. For example, in the legal domain, LLMs have successfully passed the bar exam (Katz et al. 2024) and demonstrated high accuracy in Korean legal article prediction (Hwang et al. 2022). However, in medical contexts, they achieved only a 65.6% satisfaction rate for emergency treatment guidelines, highlighting their limited applicability in urgent situations (Birkun and Gautam 2023).

Similarly, the use of LLMs in policing must be carefully evaluated for real-world deployment, given the risks of biased, incomplete, or incorrect information, and overconfident responses. These concerns are not merely theoret-

*Corresponding Author

ical. Figure 1 shows the risks of trusting LLMs in policing based solely on traditional evaluations. For example, LLM-based dispatch suggestions have shown regional and racial bias (Jain, Calacci, and Wilson 2024). Police work involves critical tasks such as crime prevention and public safety (National Assembly of the Republic of Korea 2022; Song 2013a), which demand strict compliance with laws and procedures (Roberts 2012). As a single misstep can have serious consequences, LLM integration must proceed with caution and robust validation.

Despite this necessity, no comprehensive studies have yet assessed the suitability and associated risks of using LLMs in police activities. As shown in Figure 1, which highlights a case where traditional evaluation methods were applied to actual LLM outputs, serious issues may arise in practice when such models are used without proper evaluation from a policing perspective. Although LLM-generated responses may not be considered entirely incorrect in legal or informational terms, relying on them without verification could lead to legal and ethical problems, such as unlawful arrests or improper evidence collection.

To address these critical issues, this study aims to establish a framework for assessing the appropriateness of LLMs in police operations and to evaluate whether widely used LLMs meet the necessary standards. Ultimately, the goal is to determine the direction for developing future police-specific LLMs. Accordingly, this study defines the following two key research questions:

- **RQ1**: What is an appropriate evaluation framework for LLMs in police operations that integrates domain-specific metrics and evaluation datasets?
- **RQ2**: To what extent can LLMs effectively answer inquiries about police work?

Based on the above research questions, this study identifies two major limitations in existing evaluations of LLMs for police work: (1) the lack of flexible scenario design methods and corresponding evaluation datasets tailored to real-world police response situations, and (2) the absence of a comprehensive metric framework for assessing LLM performance on police-specific tasks.

To address these gaps, we propose a specialized evaluation framework, **PAS** (Police Action Scenarios). **PAS** consists of five key stages: defining real-world police action scenarios, constructing expert reference answers for each scenario, generating LLM responses based on the scenarios, extracting core evaluation metrics, and interpreting results using LLM-based judges. The framework is designed to be flexibly adapted across diverse missions and operational contexts within the policing domain.

To validate the practical utility of **PAS**, we implemented it in the context of a *Police Readiness through Operational Reasoning* scenario. We curated and refined over 8,000 official Korean police documents and constructed an evaluation dataset to benchmark the performance of commercial LLMs. Through the experiment, we identified five core evaluation metrics that are closely aligned with key indicators of real-world police performance. The results revealed that LLM-generated responses significantly underperformed expectations. These findings provide actionable insights for improving the future applicability of LLMs in police operations.

The contributions of our work are articulated as follows:

- We propose **PAS**, a scalable evaluation framework tailored to police scenarios, covering various real-world policing scenario design, expert references, LLM responses, metric extraction, and judgment.
- We construct a curated corpus of 8,000+ Korean police documents and build an evaluation dataset for real-world LLM assessment. Field police experts participated in validating core evaluation metrics, highlighting the need for expert-in-the-loop supervision in developing reliable evaluation frameworks across diverse policing contexts.
- Our evaluation reveals that commercial LLMs underperform on key police metrics, highlighting the gap between general capabilities and domain-specific requirements.

Related Works

LLM Evaluation Methods

The evaluation of Large Language Models (LLMs) has evolved through two main approaches: multiple-choice QA and open-ended QA.(Myrzakhan, Bsharat, and Shen 2024)

Multiple-choice QA has demonstrated remarkable effectiveness in assessing professional knowledge, with LLMs showing impressive performance across various professional certification exams including bar examination, medical licensing tests, and CPA evaluations(Katz et al. 2024; Kung et al. 2023; Bommarito et al. 2023).

In parallel, open-ended QA assessment have undergone significant development. While traditional metrics like BLEU and ROUGE based on text similarity remain valid in certain contexts, they have been complemented by more sophisticated approaches that encompass multiple dimensions such as consistency, logical reasoning, and factual accuracy(Kamalloo et al. 2023). While such multi-dimensional analysis traditionally relied on human evaluators, recent studies have shown that LLMs can effectively serve as automated judges, providing a scalable solution while maintaining quality standards (Chiang and Lee 2023).

Challenges in Domain-Specific LLM Evaluation

The application of these evaluation methodologies in professional domains has revealed significant limitations. Multiple-choice QA, despite its effectiveness in standardized tests, fails to capture practical competence in real-world scenarios(Li et al. 2024). In healthcare, LLMs achieved only 65.6% satisfactory responses when evaluating emergency treatment guidelines (Birkun and Gautam 2023). Similarly, in the legal domain, while LLMs demonstrate impressive results in structured evaluations(Hwang et al. 2022), their performance significantly drops when faced with real-world police scenarios. Additional issues include potential biases in their decision-making processes(Jain, Calacci, and Wilson 2024). For open-ended QA, traditional metrics like BLEU and ROUGE are insufficient for evaluating domain-specific responses. Even with reference answers available,

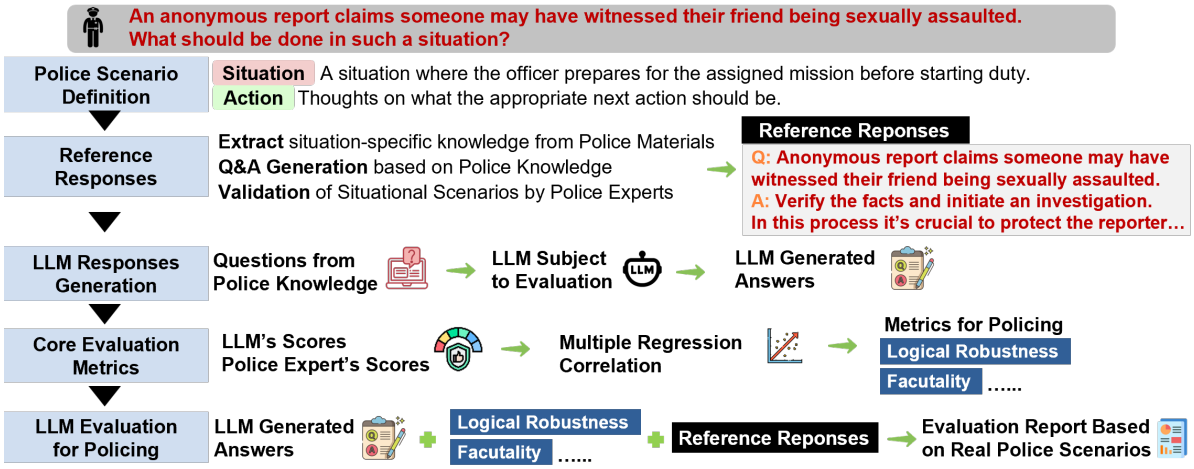

Figure 2: An overview of the **PAS**. Unlike previous benchmark studies that primarily focused on accuracy in legal matching or crime classification (Kim et al. 2024; Eun-Jung Kwon and Byon 2024; Hwang et al. 2022; Baek et al. 2021), this study adopts a Police Action Scenario-based framework to evaluate LLMs. By simulating real-world police situations, the framework generates both LLM responses and expert reference answers. Furthermore, it designs domain-specific evaluation metrics to comprehensively assess the LLM's situational applicability in policing contexts.

high BLEU or ROUGE scores may not indicate better responses - in fact, responses with high textual similarity might contradict the intended meaning or provide inappropriate guidance in professional contexts(Xu et al. 2024). While multi-dimensional evaluation approaches offer a potential solution, the lack of specialized evaluation datasets and metrics remains a significant challenge. This highlights a significant research gap in domain-specific LLM evaluation, particularly in police operations, where responses must align with operational procedures and regulatory requirements. We address this gap by proposing a specialized evaluation framework tailored to this domain.

Framework for Evaluating LLMs in Policing

Real world policing requires complex judgment and context specific actions, making scenario-based evaluation essential for assessing LLMs. However, no existing research provides a policing specific framework, and relying solely on standard metrics may lead to critical mistakes.

To address this, we propose a **PAS**, an LLM evaluation framework based on **Police Action Scenarios**. The **PAS** is constructed as a five-stage evaluation framework for policing tasks, formally expressed as $E_{\text{police}} = f(S, R, G, M, P)$. Each stage has been designed to ensure consistent applicability across varying police scenarios.

- **Policing Scenario Definition**(S) : Situation-based task.
- **Reference Responses**(R) : Construction of golden answers through participation of police experts.
- **Response Generation**(G) : LLM outputs on scenarios.
- **Core Evaluation Metrics**(M) : Selection of policing-specific metrics and evaluation methodology design.
- **LLM Performance Evaluation for Policing**(P) : Comprehensive evaluation of LLM suitability.

The **PAS** framework can be formally represented as:

$$\text{PAS} : S \xrightarrow{\text{Experts}} R, S \xrightarrow{\text{LLMs}} G \xrightarrow{(M,R)} P \quad (1)$$

Step 1: Policing Scenario Definition(S) We proposes a policing scenario-based task design to evaluate LLMs acting in the role of police officers. The objective is to assess the model's reasoning and decision-making in realistic policing contexts. We define scenarios along two key dimensions:

- **Situation (S)**: The operational state in which an officer is placed. This includes the officer's department, role, mission, and the type of incident encountered.
- **Action (A)**: The cognitive or behavioral output required in response to the situation, such as legal judgment, case classification, report generation, or citizen interaction.

This formulation is based on core principles. First, legal definitions of police duties, such as crime prevention, investigation, and public order, support the classification of the situation S (Republic of Korea 2017; Song 2013b). Policing models, including traditional and intelligence led approaches, inform how officers act across contexts (Pereira, Rosado, and Lopes 2021), shaping the definition of the action component A. Since police work relies on legal reasoning and situational judgment (Roberts 2012; Clark 2012), LLM evaluations must reflect this cognitive aspect.

Step 2: Reference Responses (R) To ensure reliable LLM evaluation for Police Action Scenarios, gold-standard reference answers are essential, as no such datasets currently exist in policing. These are built using input from local police experts and training materials, with real case data reviewed and filtered for high-reliability responses. This ensures both jurisdictional relevance and adaptability, providing strong benchmarks and resources for future LLM alignment.

Step 3: LLM Response Generation (G) In this stage, LLM responses are generated from the defined Police Action Scenarios, providing empirical insight into how well LLMs mirror police reasoning. These outputs help assess the need for police-specific LLMs. Given the complexity of policing tasks, this step also lays the groundwork for future comparisons using specialized evaluation metrics. Given a **Police Action Scenario** S, the LLM generates responses $G = \{g_1, \ldots, g_n\}$ that approximate police reasoning and judgment. Formally, $G = \text{LLM}(S)$, where the model maps scenarios to outputs under constraints defined by S.

Step 4: Core Evaluation Metrics (M) To evaluate LLMs in policing, proper metrics are essential for assessing response quality. Existing LLM metrics offer useful foundations but must be adapted to the unique demands of police work. For example, policing tasks demand more than accuracy-they require situational fit, legal-procedural alignment, and judgment rationality. Furthermore, questions remain whether conventional metrics alone can fully capture the quality of responses in police-related tasks.

Therefore, this stage defines the systematic process for selecting the final Core Metrics as follows:

- First, candidate metrics are drawn from existing LLM evaluation studies and tailored to fit policing scenarios.
- Next, the adapted metrics are applied to LLM responses across scenarios, evaluated by both an automated LLM evaluator and human police experts.
- Finally, a two-stage filtering process refines the metric set: (1) multiple regression identifies metrics that significantly predict response quality, and (2) correlation analysis validates these, keeping only those positively aligned with expert judgments.

Step 5: LLM Performance Evaluation (P) The final stage conducts a comprehensive evaluation using core metrics adapted from existing LLM research to fit policing needs. Key indicators include expert-aligned overall quality, legal and procedural consistency, and real-world applicability. This structured assessment not only validates the LLM's utility in police scenarios but also informs future development and policy for policing-focused LLM deployment.

The final evaluation integrates metrics into an overall assessment function $P = h(G, R, M)$, where h measures the alignment between generated responses G, reference responses R, and metrics M. This produces a multi-dimensional score reflecting LLM performance in policing scenarios, offering both selection benchmarks and diagnostic insights into alignment with policing standards.

Logical Thinking	**Background Knowledge**
Logical Robustness	Factuality
Logical Correctness	Groundness
Logical Efficiency	Commonsense Understanding
	Numerical Sensitivity
Problem Handling	**User Alignment**
Comprehension	Readability
Insightfulness	Conciseness
Completeness	Harmlessness
Metacognition	Logical Explanation

Table 1: Candidate Metrics for Model Evaluation. See supplements for definitions.

Category	**Item**	**Count**
Area	Cyber & Economic Crimes	17
	Violent & Serious Crimes	14
	General Investigation & Procedure	14
	Public Safety	9
	Special Crimes Investigation	4
	Traffic	4
	Victim Protection & Human Rights	3
	112 Emergency Calls	3
	Forensic Science & Evidence Analysis	2
	Crime Prevention	2
	Security	2
	Police Organization & Operations	1
Process	Investigation	51
	Initial Response & Suppression	13
	General	9
	Post-Incident Protection & Management	1
	Internal	1
Type	Function-Specific Manuals	41
	Investigation Checkpoints	20
	Criminal Law Knowledge	6
	Investigation Techniques	4
	Investigation Reporting Guidelines	4

Table 2: Statistics of Police Manuals by Category

Experiment

Applying PAS to Real Police Data

Simulating Operational Readiness In this study, we selected the scenario of *Police Readiness through Operational Reasoning*, which reflects a common form of mental training that police officers engage in before duty. The **Situation** represents the preparatory phase prior to starting work, and the **Action** involves mentally simulating responses to anticipated incidents. This practice is rooted in the daily routine of officers who regularly face high-risk situations, enabling them to respond safely and effectively. For example, a patrol officer might think, "It's the weekend and there could be more domestic violence calls. What's the protocol again?" Such mental rehearsals often include consulting manuals or past reports to plan appropriate responses in advance.

Reference Answers from Police Manuals To generate reference answers for the *Police Readiness through Operational Reasoning*, we used 1,602 official police manuals from the Korean National Police Agency. These documents span various domains, including investigation, law enforcement, traffic control, and emergency response, and contain legal references, procedural guidelines, and training materials. Most were in PDF format with 3 to 4 levels of hierarchy, requiring preprocessing for LLM input. We segmented them by section headers and reformatted the content into a standardized {title, content, question} structure. With expert guidance, we filtered for quality and compiled 8,348 curated

entries. From this set, we created 75 question and answer pairs, as shown in Table 2, covering a broad range of police duties and operational contexts. Each pair was constructed by identifying a key question and locating its answer within the same entry, simulating how officers anticipate and reason through situations before beginning their shift.

LLM Response Generation We conducted zero shot experiments using the final set of 75 questions with three commercial LLMs: GPT-4, Gemini, and Claude (temperature = 0.8). Each model was given only a brief scenario describing a police officer's preparatory state before starting duty, and was instructed to respond solely from the perspective of a police officer, outlining the appropriate actions to take without access to any external knowledge or additional context.

Core Metric Alignment We then identified the optimal metrics through evaluation by experts, as outlined below.

- **Candidate Metrics Using FLASK**: We began with the 12 core metrics from the FLASK framework (Ye et al. 2024), a validated tool for evaluating LLMs in areas like logical reasoning and problem solving. We expanded this set by splitting factuality into factuality and groundness, and adding numerical sensitivity and logical explanation, inspired by Sun et al.(Sun et al. 2024). In total, we defined 15 candidate metrics to assess both general LLM abilities and policing-specific needs. Definitions are provided in Table 1.
- **LLM-as-a-Judge & Police Expert Evaluation**: To identify key metrics for policing-specific applications, we evaluated 225 responses (75 questions across 3 LLMs) and reference answers using the 15 candidate metrics and an overall quality score on a 5 point Likert scale. For the automated assessments, the temperature of the LLM-as-a-Judge was set to 0.8. Both LLM-as-a-Judge and police experts assessed the responses, combining three automated evaluations with two expert reviews. This hybrid method balances prior findings on LLM judge reliability (Hada et al. 2024; Chiang and yi Lee 2023) with caution around automating policing-specific evaluation. To reduce this uncertainty, we used structured prompts based on prior LLM evaluation frameworks (Ye et al. 2024; Zheng et al. 2023). Two strategies were employed to enhance reliability. First, each prompt included an expert written reference response alongside the target response, which improves alignment with human judgments (Krumdick et al. 2025) and helps detect domain-specific errors (Ryu et al. 2023). Second, prompts required judges to explain their reasoning before assigning a score to increase agreement with human evaluators (Chiang and Lee 2023), especially when guided by a scoring rubric (Pires, Junior, and Nogueira 2025).
- **Key Metrics Determination**: To identify the most meaningful evaluation metrics, we implemented a two-stage filtering process. First, a stepwise regression method was applied to determine which of the 15 candidate metrics were significant predictors of the overall quality score. Only metrics with a significance level of $p < .05$ advanced to the second stage. Next, these selected candidates were validated against the judgments of human police experts using Spearman's rank correlation (ρ). A metric was retained for the final set only if it also showed a statistically significant positive correlation with the expert evaluations ($p < .05$). This rigorous process was designed to ensure that our final metrics possess both statistical predictive power and domain relevance.

	Regression			Correlation		Final
Metric	β	**SE**	***p*-val**	ρ	***p*-val**	**Sel.**
L. Correctness	0.116	0.049	.018	0.560	<.001	✓
Completeness	0.141	0.042	.001	0.409	<.001	✓
Factuality	0.247	0.048	<.001	0.330	.002	✓
L. Efficiency	0.137	0.048	.004	0.330	.002	✓
L. Robustness	0.167	0.052	.002	0.315	.003	✓
L. Explanation	0.108	0.036	.003	0.203	.055	×

Table 3: Two-stage analysis for key metric selection. Metrics were selected if they were both statistically significant predictors in multiple regression ($p < .05$) and showed a significant positive correlation with expert judgments ($p < .05$).

Experiment Results

RQ1: What Is an Appropriate Evaluation Framework for LLMs in Police Operations That Integrates Domain-Specific Metrics and Evaluation Datasets?

Two-Stage Analysis for Key Metric Selection A multiple regression analysis was conducted to determine the most significant metrics for evaluating LLM-generated responses in police operations. Among the 15 candidate metrics, six were found to be statistically significant in predicting response quality. The model exhibited a strong fit ($R^2 = 0.856$, $F = 297.457$, $n = 300$, $p < 0.001$), suggesting that these criteria are essential for the effective evaluation of LLM-generated responses in police operations.

Next, these candidates were validated against the judgments of human police experts using Spearman's rank correlation. As shown in Table 3, five of the six metrics demonstrated a statistically significant positive correlation with the expert evaluations (p $<.05$). However, we note that the correlation coefficients themselves were moderate, ranging from ρ=0.315 to ρ=0.560, which is consistent with known challenges in automated evaluation of specialized domains (Szymanski et al. 2025). This indicates that while the LLM-as-a-judge scores are directionally aligned with expert opinions, they are not a perfect substitute.

The final key metrics are: Logical Correctness, Completeness, Factuality, Logical Efficiency, and Logical Robustness.

RQ2: To What Extent Can LLMs Effectively Answer Inquiries about Police Work?

Reference responses demonstrate high performance across evaluation metrics The evaluation of the reference answers shows strong overall performance (3.88/5.0), with particularly high scores in factuality (4.15), logical correctness (4.14), and harmlessness (4.11). These results validate

Metric	All	Claude	Gemini	GPT-4
L. Robustness	-0.93^{***}	-0.97^{***}	-0.72^{***}	-1.09^{***}
L. Correctness	-1.32^{***}	-1.34^{***}	-1.10^{***}	-1.52^{***}
L. Efficiency	-1.09^{***}	-1.08^{***}	-0.97^{***}	-1.22^{***}
Factuality	-1.38^{***}	-1.43^{***}	-1.16^{***}	-1.54^{***}
Groundness	-1.41^{***}	-1.48^{***}	-1.17^{***}	-1.57^{***}
Common. Und.	-1.05^{***}	-1.03^{***}	-0.86^{***}	-1.27^{***}
Num. Sensitivity	-1.38^{***}	-1.39^{***}	-1.18^{***}	-1.56^{***}
Comprehension	-1.06^{***}	-1.11^{***}	-0.84^{***}	-1.24^{***}
Insightfulness	-0.38^{*}	-0.58^{***}	-0.01	-0.54^{***}
Completeness	-0.86^{***}	-1.03^{***}	-0.51^{***}	-1.04^{***}
Metacognition	-0.31^{**}	-0.35^{**}	-0.02	-0.57^{***}
Readability	-0.13	-0.21^{*}	0.04	-0.23
Conciseness	-0.37^{**}	-0.18	-0.40^{**}	-0.52^{***}
Harmlessness	-0.20^{*}	-0.29^{**}	-0.01	-0.32^{***}
L. Explanation	-0.55^{***}	-0.72^{***}	-0.24	-0.70^{***}
Overall	-1.02^{***}	-1.04^{***}	-0.81^{***}	-1.19^{***}

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Table 4: T-test Results on Performance Gaps Between LLMs and Reference Answers. Key metrics identified through multiple regression analysis are in **bold**, and underlined values represent the five metrics where each LLM shows the greatest deviation from the reference answers, primarily in Logical Thinking and Background Knowledge.

the quality of our manual-based reference answers, which serve as the standard for assessing LLMs' performance.

Identified Key Metrics Reveal Critical Gaps in LLM Performance Analysis of the identified key metrics revealed significant deficiencies in LLM performance. As shown in Table 4, the average performance across Claude, GPT-4, and Gemini was significantly lower than the reference answers across these critical metrics ($p<.05$), with an average skill difference of -1.115 in these key metrics (compared to -0.828 for all evaluation skills).

Model-wise analysis reinforced these findings, with all model-metric combinations showing significantly lower performance than reference answers ($p < .05$), except for Gemini's logical explanation. The performance gap was particularly pronounced in these key metrics. Specifically, Gemini exhibited a difference of -0.893, Claude -1.169, and GPT -1.283, which were notably larger than their respective all-metric average differences of -0.611, -0.880, and -0.994.

This performance gap extends to all evaluation metrics. A comparison with reference answers revealed substantial differences across all models. Specifically, the average difference in overall scores was -1.015. Statistical analysis using t-tests confirmed that, except for readability, all skills exhibited significantly lower scores compared to reference answers ($p < .05$). These findings reinforce the importance of the identified key metrics and underscore the significant limitations of current LLMs in effectively handling police-work-related queries.

Category	**Reference**	**GPT-4**	**Gemini**	**Claude**
Overall Score	3.88	2.69	3.06	2.83
Skill Score	3.72	2.73	3.11	2.84
Factuality	4.15	2.60	2.98	2.72
L. Correctness	4.14	2.61	3.04	2.80
Harmlessness	4.11	3.80	4.10	3.82
Common. Und.	4.00	2.73	3.14	2.97
Groundness	3.91	2.33	2.73	2.42
Comprehension	3.87	2.63	3.03	2.76
L. Efficiency	3.84	2.63	2.87	2.77
Completeness	3.81	2.76	3.30	2.78
L. Robustness	3.81	2.72	3.09	2.84
Readability	3.80	3.59	3.84	3.57
Num. Sensitivity	3.69	2.12	2.51	2.30
L. Explanation	3.60	2.90	3.37	2.89
Insightfulness	3.47	2.92	3.45	2.88
Metacognition	2.85	2.28	2.83	2.50
Conciseness	2.76	2.25	2.36	2.58

Table 5: Overall scores of LLMs and comparison of performance across metrics.

LLMs are weak in providing high-quality responses to more specialized police knowledge Table 5 compares LLM-generated responses with reference answers, showing that reference answers consistently achieve higher scores in overall performance (3.88) and skill-specific metrics (3.72). As illustrated in Table 5 and 6, this superiority extends across all skill categories and work domains, reinforcing the assumption that police manual-based responses exhibit higher quality than LLM-generated ones. The findings further highlight that LLMs do not consistently produce high-quality responses in all aspects of police work.

A detailed analysis of skill-specific scores indicates that LLMs underperform primarily due to a "lack of expertise in police work." Reference answers score highest in factuality (4.15), logical correctness (4.14), and harmlessness (4.11), demonstrating strong domain-specific knowledge. In contrast, LLMs perform best in harmlessness (3.91), readability (3.67), and insightfulness (3.09), emphasizing their strength in text coherence over factual accuracy.

This gap arises because reference answers originate from police manuals rather than being optimized for LLM prompts. While LLMs generate well-structured text, they lack specificity and factuality in police-related contexts.

Regarding police work areas, answer models handle broad topics but struggle with specialized subjects. Table 6 shows that reference answers score highest in Crime Prevention (4.11) and Violent and Major Crimes (4.05). Among answer models, GPT-4 excels in crime prevention (3.33) and victim protection (2.89), while Gemini and Claude achieve their highest scores in police organization and operations. However, all models consistently score low in 112 emergency calls (2.37), traffic (2.55), and security (2.61), highlighting persistent limitations across these domains.

Work Domain	Ref.	GPT-4	Gemini	Claude
Crime Prevention	4.11	3.22	3.33	3.11
Violence & Major Crimes	4.05	2.86	3.21	3.02
Forensic Science	4.00	2.83	3.50	3.00
Public Safety	4.00	2.63	2.85	2.89
Victim Protection	4.00	2.89	3.33	3.22
Traffic	3.92	2.50	2.67	2.50
General Investigation	3.90	2.64	2.98	2.69
Security	3.83	2.33	2.67	2.83
Cyber & Economic Crimes	3.78	2.63	3.10	2.98
112 Report	3.33	2.33	2.89	1.89
Special Crime Investigation	3.33	2.67	3.25	2.33

Table 6: Comparison of performance across work domains.

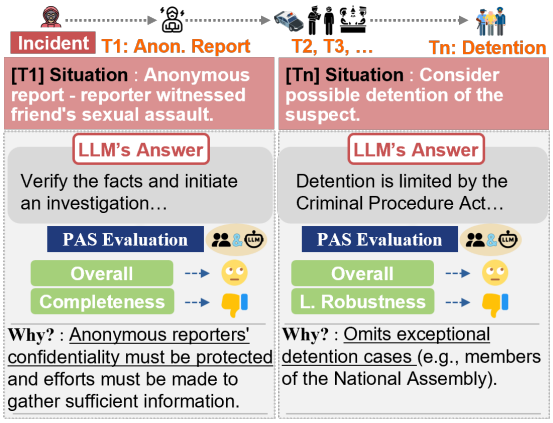

Figure 3: **PAS** Application in Real Police Work Sequences

Discussion

PAS Usefulness for LLM Evaluation in Policing

We explore how applying the **PAS** framework enables a more realistic evaluation of LLMs in policing, compared to traditional methods. Figure 3 presents a case from our experimental scenario, highlighting a typical sequence of tasks drawn from the *Police Readiness through Operational Reasoning*. Although the LLM responses to these questions contain no factual errors and may receive high scores under standard evaluations, they fall short in delivering manual based procedures, practical knowledge, or warnings about legal risks such as unlawful arrest. This suggests that relying on such responses could raise the risk of misconduct. By using **PAS**, we gain clearer insight into LLM limitations in police contexts and can better identify the outputs that need improvement for practical deployment.

Generalizability and Practical Pathways

A critical challenge for deploying AI in policing is its adaptation to diverse legal and cultural contexts. The PAS framework directly addresses this issue of generalizability through its Reference Responses (R) component. Rather than enforcing a single standard, our framework is designed to be flexibly adapted by having local police experts build the Reference Responses aligned with their own jurisdiction's laws and operational environment. For instance, a reference response based on the Korean Criminal Procedure Act can be replaced with one tailored to U.S. procedures, allowing the same framework structure-with Step 4 enabling metric adaptation as needed for each context-to accurately evaluate LLM performance in a different policing context. This scalable approach enables agencies to measure and improve LLM performance within their unique settings, providing a significant advantage for safe, real-world integration.

Implications of Moderate Correlation: The Need for an Expert-in-the-Loop Framework

The moderate correlation between LLM judges and expert evaluations reveals the limitations of LLM-as-a-judge, reinforcing the importance of our framework's multi-stage validation approach.Recent studies reveal significant limitations of LLM-as-a-judge in expert knowledge domains, with pairwise preference agreement between LLM judges and Subject Matter Experts as low as 60-64% in specialized healthcare fields (Szymanski et al. 2025). LLMs often align more closely with lay-user preferences than with domain expert criteria (Bavaresco et al. 2025), a challenge particularly pronounced in high-stakes policing where training data is scarce and expert judgment is critical. Unlike previous approaches that rely solely on automated metrics, **PAS** incorporates domain expert validation at key stages, addressing these fundamental limitations of LLM-as-a-judge in specialized fields.

Conclusion

This study proposes a comprehensive evaluation framework, **PAS**, developed to address the complexities of police operations. We defined highly realistic police action scenarios and designed a generation method to elicit LLM outputs under those contexts, while also constructing golden answers that enable expert evaluation. As a result, we built QA datasets from over 8,000 police manuals and identified five key indicators—Logical Correctness, Completeness, Factuality, Logical Efficiency, and Logical Robustness—as strong predictors of response quality. Our results demonstrate that commercial LLMs consistently underperform in tasks requiring factual and procedural precision. These findings emphasize the limitations of existing LLMs in specialized domains like policing and the need for research to align outputs with real-world police standards. The **PAS** not only offers a replicable evaluation benchmark but also provides a practical method applicable to other professional fields. As LLMs become more integrated into public safety systems, such frameworks are vital to ensure reliable and lawful use.

Our experiments also highlight the mixed potential of LLM-as-a-Judge. While scores moderately align with expert assessments, they cannot fully replace human judgment in high-stakes contexts. Moreover, since our evaluation was based on Korean manuals, broader applicability may be limited. Future work should expand **PAS** with real-time data, jurisdictional diversity, and refined evaluation strategies.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(Ministry of Science and ICT) RS-2025-16064585.

References

Adams, I. T. 2024. Large Language Models and Artificial Intelligence for Police Report Writing. *CrimRxiv*.

Baek, M.-S.; Park, W.; Park, J.; Jang, K.-H.; and Lee, Y.-T. 2021. Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation. *IEEE Access*, 9: 131906–131915.

Bavaresco, A.; Bernardi, R.; Bertolazzi, L.; Elliott, D.; Fernández, R.; Gatt, A.; Ghaleb, E.; Giulianelli, M.; Hanna, M.; Koller, A.; Martins, A. F. T.; Mondorf, P.; Neplenbroek, V.; Pezzelle, S.; Plank, B.; Schlangen, D.; Suglia, A.; Surikuchi, A. K.; Takmaz, E.; and Testoni, A. 2025. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. arXiv:2406.18403.

Birkun, A. A.; and Gautam, A. 2023. Large Language Model (LLM)-powered chatbots fail to generate guideline-consistent content on resuscitation and may provide potentially harmful advice. *Prehospital and Disaster Medicine*, 38(6): 757–763.

Bommarito, J.; Bommarito, M.; Katz, D. M.; and Katz, J. 2023. GPT as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities. arXiv:2301.04408.

Chiang, C.-H.; and Lee, H.-y. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8928–8942.

Chiang, C.-H.; and yi Lee, H. 2023. Can Large Language Models Be an Alternative to Human Evaluations? arXiv:2305.01937.

Clark, D. 2012. Covert surveillance and informer handling. In *Handbook of criminal investigation*, 426–449. Willan.

Eun-Jung Kwon, M. L., Hyunho Park; and Byon, S. 2024. Validation of Training Data for AI-Based 112 Emergency Reporting. *Proceedings of Symposium of the Korean Institute of communications and Information Sciences*, 904–905.

Fei, Z.; Shen, X.; Zhu, D.; Zhou, F.; Han, Z.; Zhang, S.; Chen, K.; Shen, Z.; and Ge, J. 2023. LawBench: Benchmarking Legal Knowledge of Large Language Models. arXiv:2309.16289.

Hada, R.; Gumma, V.; de Wynter, A.; Diddee, H.; Ahmed, M.; Choudhury, M.; Bali, K.; and Sitaram, S. 2024. Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation? arXiv:2309.07462.

Halford, E.; and Webster, A. 2024. Using chat GPT to evaluate police threats, risk and harm. *International Journal of Law, Crime and Justice*, 78: 100686.

Hwang, W.; Lee, D.; Cho, K.; Lee, H.; and Seo, M. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35: 32537–32551.

Jain, S.; Calacci, D.; and Wilson, A. 2024. As an AI Language Model, "Yes I Would Recommend Calling the Police": Norm Inconsistency in LLM Decision-Making. arXiv:2405.14812.

Jamal, S.; and Wimmer, H. 2023. An Improved Transformer-based Model for Detecting Phishing, Spam, and Ham: A Large Language Model Approach. arXiv:2311.04913.

Kamalloo, E.; Dziri, N.; Clarke, C.; and Rafiei, D. 2023. Evaluating Open-Domain Question Answering in the Era of Large Language Models. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Katz, D. M.; Bommarito, M. J.; Gao, S.; and Arredondo, P. 2024. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270): 20230254.

Kim, H.; Kim, D.; Lee, J.; Yoon, C.; Choi, D.; Gim, M.; and Kang, J. 2024. LAPIS: language model-augmented police investigation system. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4637–4644.

Kim, H.-D.; and Lim, H. 2022. A named entity recognition model in criminal investigation domain using pretrained language model. *Journal of the Korea Convergence Society*, 13(2): 13–20.

Krumdick, M.; Lovering, C.; Reddy, V.; Ebner, S.; and Tanner, C. 2025. No Free Labels: Limitations of LLM-as-a-Judge Without Human Grounding. arXiv:2503.05061.

Kung, T. H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. 2023. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. *PLoS digital health*, 2(2): e0000198.

Li, W.; Li, L.; Xiang, T.; Liu, X.; Deng, W.; and Garcia, N. 2024. Can Multiple-choice Questions Really Be Useful in Detecting the Abilities of LLMs? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2819–2834.

Liu, F.; Zhou, H.; Hua, Y.; Rohanian, O.; Clifton, L.; and Clifton, D. A. 2024a. Large Language Models in Healthcare: A Comprehensive Benchmark. *medRxiv*.

Liu, M.; Hu, W.; Ding, J.; Xu, J.; Li, X.; Zhu, L.; Bai, Z.; Shi, X.; Wang, B.; Song, H.; et al. 2024b. Medbench: A comprehensive, standardized, and reliable benchmarking system for evaluating chinese medical large language models. *Big Data Mining and Analytics*, 7(4): 1116–1128.

Myrzakhan, A.; Bsharat, S. M.; and Shen, Z. 2024. Open-LLM-Leaderboard: From Multi-choice to Open-style Questions for LLMs Evaluation, Benchmark, and Arena. arXiv:2406.07545.

National Assembly of the Republic of Korea. 2022. ACT ON THE ORGANIZATION AND OPERATION OF NATIONAL POLICE AND AUTONOMOUS POLICE. Article 3 (Duties of Police).

Pereira, A. R. P.; Rosado, D. P.; and Lopes, H. S. C. 2021. From the Traditional Police Model to Intelligence-Led Policing Model: Comparative Study. In *Information and Knowledge in Internet of Things*, 457–473. Springer.

Pires, R.; Junior, R. M.; and Nogueira, R. 2025. Automatic Legal Writing Evaluation of LLMs. arXiv:2504.21202.

Republic of Korea. 2017. Act on the Persons Performing the Duties of Judicial Police Officers and the Scope of Their Duties. Available at: https://www.law.go.kr/. [Enforcement Date: Dec 19, 2017.] [Act No.15253, Partial Amendment].

Roberts, P. 2012. Law and criminal investigation. In *Handbook of criminal investigation*, 92–145. Willan.

Ryu, C.; Lee, S.; Pang, S.; Choi, C.; Choi, H.; Min, M.; and Sohn, J.-Y. 2023. Retrieval-based evaluation for LLMs: a case study in Korean legal QA. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, 132–137.

Sarzaeim, P.; Mahmoud, Q. H.; and Azim, A. 2024. A framework for LLM-assisted smart policing system. *IEEE Access*, 12: 74915–74929.

Song, S. 2013a. On the Legal Basement and Limits of the Police Power in the USA -An Essay to Interpret Police Enforcement Law in Principle. *Journal of hongik law review*, 14(1): 669–699.

Song, S. K. 2013b. On the Legal Basement and Limits of the Police Power in the USA: An Essay to Interpret Police Enforcement Law in Principle. *The Law Research Institute of Hongik University*, 14(1): 669–699.

Stotland, E. 1991. The effects of police work and professional relationships on health. *Journal of Criminal Justice*, 19(4): 371–379.

Sun, Z.; Shen, Y.; Zhou, Q.; Zhang, H.; Chen, Z.; Cox, D.; Yang, Y.; and Gan, C. 2024. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36.

Szymanski, A.; Ziems, N.; Eicher-Miller, H. A.; Li, T. J.-J.; Jiang, M.; and Metoyer, R. A. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 952–966.

Tan, Y. S.; Zalzuli, A. D.; Ang, J.; Ho, H. F.; and Tan, C. 2022. Understanding the workload of police investigators: a human factors approach. *Journal of police and criminal psychology*, 37(2): 447–456.

Tong, X.; Jin, B.; Lin, Z.; Wang, B.; Cheng, Q.; and Yu, T. 2024. CPSDbench: a large language model evaluation benchmark and baseline for Chinese public security domain. *International Journal of Data Science and Analytics*, 1–30.

Vila, B. 2006. Impact of long work hours on police officers and the communities they serve. *American journal of industrial medicine*, 49(11): 972–980.

Xu, S.; Wu, Z.; Zhao, H.; Shu, P.; Liu, Z.; Liao, W.; Li, S.; Sikora, A.; Liu, T.; and Li, X. 2024. Reasoning before Comparison: LLM-Enhanced Semantic Similarity Metrics for Domain Specialized Text Analysis. arXiv:2402.11398.

Ye, S.; Kim, D.; Kim, S.; Hwang, H.; Kim, S.; Jo, Y.; Thorne, J.; Kim, J.; and Seo, M. 2024. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. In *The Twelfth International Conference on Learning Representations*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.