Can Large Language Models Effectively Support Decision-Making in Sudden Emergencies?

Mengna Zhu[1*], **Jibing Wu**[1*], **Lihua Liu**[1], **Yuran Gong**[2], **Yang Hao**[1], **Fu Yachao**[1], **Mao Wang**[1†], **Lei Hou**[3†], **Juanzi Li**[3]

[1] Laboratory for Big Data and Decision, National University of Defense Technology
[2] School of Information and Software Engineering, University of Electronic Science and Technology of China
[3] Department of Computer Science and Technology, Tsinghua University
zhumengna16@nudt.edu.cn

Abstract

Emergency response is a safety-critical public governance task that demands accurate and timely decision-making based on complex event information. This process involves multiple stages, including information collection, integration, analysis, risk assessment, and decision recommendation. Existing research has predominantly concentrated on the earlier stages, while studies focusing on the decision support phase remain underexplored, primarily due to the lack of suitable datasets for reliable and compliance-aware decision-oriented modeling and evaluation. To bridge this gap, we introduce the first real-world **E**mergency **D**ecision-**M**aking dataset EDM-Bench, comprising 1,179 instances spanning diverse task formats, including judgment, choice, short-answer, and structured emergency report generation. We also construct a structured rule repository, EDM-R^2, which contains 3,406 parsed emergency regulations to enhance decision reliability. Building on these resources, we propose a rule-enhanced reasoning framework, $R^3V\text{-}EDM$, which integrates external regulatory knowledge with constrained inference mechanisms to improve both decision safety and interpretability. Extensive experiments demonstrate the inherent complexity of emergency decision-making and validate the effectiveness of our approach in enabling more reliable and trustworthy decisions.

1 Introduction

Emergencies such as *earthquakes*, *floods*, and *terrorist attacks* are among the most prototypical and high-risk scenarios in public governance. They are characterized by sudden onset, rapid development, and continuously evolving information. Effective emergency management requires rapid information analysis, precise situational assessment, and timely response formulation. Accurate decision-making is crucial to minimize human and economic losses (Zhang et al., 2018; Su et al., 2022). Efficiently understanding such complex event information and making rational, evidence-based decisions has become a critical challenge and a promising application for natural language processing (NLP).

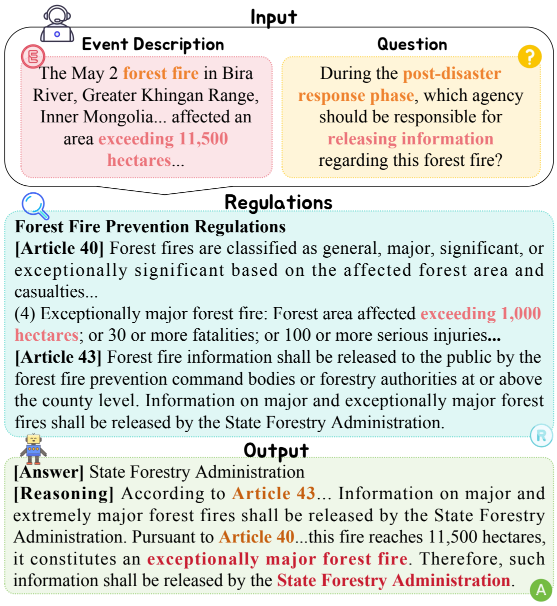

Figure 1: Illustration of reliable EDM (emergency decision-making) response generation by combining event context with relevant regulations.

Emergency response is inherently complex, often requiring integration of event-specific understanding with multiple relevant regulations to guide decision-making. For example, as shown in Figure 1, a forest fire affecting 11,500 hectares requires formulating an information release plan. According to *Article 40* of the Fire Prevention Regulations, fires impacting over 1,000 hectares are classified as exceptionally major, confirming severity. Furthermore, *Article 43* mandates that information release for such fires be handled by the State Forest Administration, defining the responsible authority.

Recent research has made significant progress in understanding emergency events through tasks

Findings of the Association for Computational Linguistics: ACL 2026, pages 36536–36558

such as event timeline summarization (Faghihi et al., 2022), event extraction (Yan et al., 2022), and knowledge graph construction (Mu et al., 2024). These efforts have laid a solid foundation for emergency comprehension. However, the effective integration and utilization of this knowledge to support practical emergency decision-making remain largely underexplored. While LLMs offer powerful language understanding and generation capabilities, opening new possibilities across NLP tasks, the complexity of emergency scenarios presents significant challenges. Sole reliance on generative LLMs still falls short of meeting the stringent reliability and safety requirements essential for high-stakes applications. Key challenges in emergency decision-making research include: 1) **Data scarcity and complexity:** High-quality, comprehensive datasets are scarce due to fragmented raw information and limited public availability (Chen et al., 2024; Hu et al., 2007); 2) **Task complexity and lack of unified definitions:** Emergency decision-making demands multi-level, multi-step reasoning over complex scenarios, yet lacks widely accepted task formulations and evaluations; 3) **Knowledge dependency and reasoning difficulties:** Accurate decisions require comprehensive regulatory and contingency knowledge, beyond models' internal capacities; 4) **Strict safety and interpretability requirements:** Safety-critical scenarios require transparency and high reliability, which current methods often fail to guarantee.

To address these challenges, we formalize the emergency decision-making task across diverse formats, including judgment, multiple-choice, multiple-answer, short-answer questions, and structured emergency response report generation. This comprehensive coverage spans fact verification, rule selection, integrative analysis, and report synthesis. Incorporating external knowledge such as regulations and emergency plans ensures safer, more controllable, and compliant model outputs.

Main contributions are summarized as follows:[1]

1. Formalized the emergency decision-making task across diverse formats and introduced EDM-Bench, the first real-world dataset with 1,179 samples that integrate authentic event descriptions, regulatory texts, and actual response reports, ensuring high quality and practical relevance;

2. Developed a structured rule repository EDM-R^2 comprising 3,406 hierarchical emergency regulations and contingency plans, enabling efficient retrieval and reasoning, significantly enhancing the model's external knowledge support;

3. Proposed a rule-enhanced reasoning framework $R^3V\text{-}EDM$, that integrates rule retrieval, reasoning constraints, evidence chain generation, as well as verification and refinement steps to enhance decision reliability and trustworthiness;

4. Defined tailored evaluation metrics for each task format and conducted comprehensive experiments benchmarking advanced LLMs, highlighting task complexity and effectiveness of $R^3V\text{-}EDM$.

2 Task Formulation

We define emergency response as an event-driven decision-making task requiring situational understanding and regulatory reasoning. Given an event description E, the model must perform reasoning grounded in context and regulations to produce compliant decision outputs. Our evaluation spans five formats: True or False (TF), Multiple Choice Question (MCQ), Multiple Answer Question (MAQ), Short Answer Generation (SAG), and Report Generation (RG), covering scenarios from factual judgment to complex action planning.

Formally, the task is modeled as $f_\theta(x) = (A, T)$, where the input $x = (E, I, O)$. Here, E provides factual information on the event's evolution; I specifies the decision task (e.g., response level determination or report modules); and $O = \{o_1, \ldots, o_k\}$ contains candidate answers for selection tasks ($k = 2$ for TF; $k = 4$ for MCQ; $k = 5$ for MAQ), while $O = \varnothing$ for SAG and RG.The output consists of an answer A and a reasoning trace T. For selection tasks, $A \in O$, where TF and MCQ have only one correct option, while MAQ has two or more correct options. For SAG, A is a concise natural language response; for RG, A is a standardized report covering assessment and measures. T provides the interpretative reasoning process based on E and relevant regulations, ensuring the traceability of the final decision.

3 Data Construction

In this section, to comprehensively investigate the performance of LLMs in emergency decision-making tasks, we construct the EDM-Bench dataset following a predefined taxonomy derived from relevant regulations and collected corpora. The taxonomy is defined according to the *Chinese Emergency*

[1]The data are in Chinese, with examples translated into English for clarity in this paper.

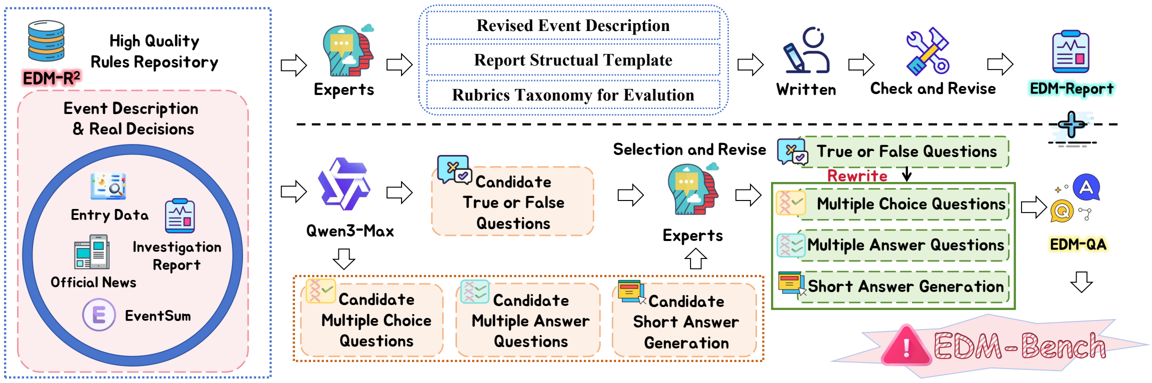

Figure 2: Construction Method of EDM-Bench, which consists of EDM-QA and EDM-Report.

Response Plans for Emergencies and covers four major categories, natural disasters, accident incidents, public health emergencies, and social security incidents, encompassing 11 commonly occurring event types that are clearly specified in official emergency response plans and span a wide range of real-world scenarios. The dataset covers multiple task formats, including judgment, multiple-choice, short-answer, and report generation, and is divided into two main subsets, question answering data EDM-QA and report generation data EDM-Report.

The construction method of EDM-Bench is illustrated in Figure 2. To facilitate the incorporation of regulatory knowledge, we first structurally parse document-style regulations to build a rule repository EDM-R^2. For EDM-QA, candidate instances are generated by LLMs based on event descriptions and rules randomly retrieved from the rule repository, and are subsequently manually reviewed, filtered, and refined to ensure high quality. To assess the sensitivity of LLMs to question formulations and fine-grained details, we further perform question type reformulation by rewriting a selected subset of TF questions into other task formats using LLMs. For EDM-Report, key factual information is revised, and reports are manually written with reference to real emergency response reports, expert-summarized templates, and relevant regulations. After multiple rounds of cross-validation and random sampling checking, the final dataset EDM-Bench is obtained.[2]

3.1 Construction of Rule Repository EDM-R^2

Emergency response to sudden events typically involves multi-agency coordination and complex procedures, requiring precise assessment strictly based on relevant regulations. However, emergencies occur abruptly and demand rapid event evaluation and formulation of response plans. Although LLMs have demonstrated strong performance in natural language processing, they often suffer from severe hallucination issues due to insufficient domain expertise in emergency decision-making tasks, such as fabricating regulatory clauses. Accurate comprehension of regulatory details is critical in actual emergency processes, where handling a single event may require referencing multiple lengthy legal documents, but understanding long texts remains a significant challenge for LLMs (Liu et al., 2025; Bai et al., 2025; Li et al., 2024).

To enable LLMs to accurately capture and apply regulatory knowledge in decision-making, we organize raw regulatory documents into a comprehensive emergency rule repository, EDM-R^2. We collect event-specific regulations from authoritative official sources and automatically extract key elements, including *trigger conditions*, *decision authorities*, and *response actions*, using rule-based techniques (e.g., regular expression matching) to chunk and itemize lengthy documents into a structured repository. This structured formulation substantially improves the efficiency and reliability of regulatory understanding.

Data Source Based on investigation reports of real emergency events and the regulatory documents referenced therein, we collected legal and regulatory files, ordinances, and contingency plans issued by official sources such as the Ministry of Emergency Management of China, the National Laws and Regulations Database, and the Chinese Government Portal. In total, 38 documents were collected, averaging 3.5 documents per event type.

[2]The predefined taxonomy is in Appendix B. Details, prompts, quality control, and data analysis for EDM-R^2 and EDM-Bench are in Appendix C and D.

Construction Method Given the well-defined structure of regulatory documents, we segment them into chapters and clauses using regular expression matching. The segmented text is then fed into LLMs with explicit instructions to convert the content into a standardized, itemized format that is easy to search and manage. Automated annotation focuses on each regulation's *summary*, *trigger conditions*, *decision-making authorities*, and *actions*, thereby constructing the rule repository EDM-R^2.

3.2 Construction of EDM-Bench

Data Sources Following data acquisition strategies similar to related works (Ghalandari et al., 2020; Zhu et al., 2025), we first retrieved relevant emergency events from authoritative encyclopedic resources including Baidu Baike, Wikipedia, and Sogou Baike to obtain event descriptions containing overviews. Additionally, detailed event information such as official bulletins and investigation reports were collected from official news portals like China Central Television, People's Daily, and Phoenix News, as well as from the Ministry of Emergency Management (P.R China), provincial and municipal government websites, and typical case repositories, to ensure completeness and richness of the event descriptions.

EDM-QA Construction Method The construction of the question-answering(QA) dataset mainly involves a hybrid approach combining candidate QA generation by LLMs with multi-round manual filtering, verification, and correction.

Step 1: Event Description Cleaning. Raw event descriptions often contain information related to accident assessment, emergency response, and investigation, which may introduce unintended cues for model generation. We employ LLMs to clean the event descriptions, removing all content related to event evaluation and response actions, while retaining only the basic factual information.

Step 2: Candidate QA Pair Generation. To maximize the diversity of the QA data, we generate candidate QA pairs by prompting LLMs with the cleaned event descriptions and randomly retrieved rules from the constructed rule base EDM-R^2. The LLM is instructed to produce multiple candidate QA pairs with different question types, difficulty levels, and corresponding reasoning traces. Specifically, we define three difficulty levels: **1) Easy:** The answer can be directly inferred from a single rule, such as simple rule matching; **2) Medium:** The task requires combining event information with basic reasoning, such as simple calculations or selecting the appropriate rule based on event characteristics; **3) Complex:** The task involves multiple rules and multi-hop reasoning. The questions often contain mixed factual information, where most details are correct but include localized misleading elements (e.g., substituted authorities, quantities, or actions), or composite decision-making questions.

Step 3: Manual Filtering and Correction. Due to the prevalence of hallucinations in LLM-generated reasoning traces, which often results in incomplete or incorrect outputs, all candidate QA pairs undergo meticulous manual review and correction. Human intervention primarily addresses the following aspects: 1) Balancing the distribution of difficulty levels to avoid an excess of questions answerable without event-specific information; 2) Removing questions or answer options that clearly contradict common sense; 3) Verifying that decision outcomes align with real-world emergency response practices, ensuring that human experts can correctly answer the questions based on the event descriptions and regulatory texts; 4) Reviewing reasoning traces to confirm the accuracy of applied rules and the completeness and consistency of the reasoning with emergency response logic, with revisions made as necessary.

EDM-Report Construction Method Considering the high standardization of real emergency response reports in both structure and content, we design a unified report generation template by analyzing official investigation and response reports alongside emergency plans. This template constrains the content scope and expression format of generated reports and is built around core information modules common to most scenarios: basic event description, severity and risk assessment, reporting and notification, response level determination, key response measures and plans, and post-response actions. These modules consistently appear in official reports and directly reflect regulatory and procedural requirements.

To mitigate potential data leakage, we modify key factual details of real events (e.g., *time*, *location*, *casualty numbers*, and *economic losses*). Response decisions are then manually written and organized into the corresponding template fields by referring to real investigation reports and relevant regulations, resulting in reference reports with clear structure and complete information.

To facilitate an effective evaluation of the quality and reliability of emergency response decision-making reports, referring to recent works related to LLM-as-Judge (He et al., 2025; Kim, 2025; Pathak et al., 2025), we formulate targeted questions focusing on key aspects such as responsible agencies, response levels, and actions, and design corresponding rubrics (e.g., *Whether a Level III emergency response was initiated in accordance with the contingency plan in the report?*) based on the corresponding modules in the report template.

3.3 Data Analysis

We collected a total of 3,406 regulatory entries for EDM-R^2. The statistical distribution of these rules is illustrated in Figure 3 (a).

Regarding EDM-Bench, the final dataset encompasses 144 events and 1,179 data instances, comprising 1,124 QA pairs and 55 report instances. For EDM-QA, we analyzed the distribution of difficulty levels and task formats, as detailed in Figure 3 (b) and (c). In EDM-Report, the manually annotated answers vary in length from 1,118 to 6,002 characters, with a mean length of 2,827 characters. Furthermore, we annotated a total of 568 rubrics; each report contains between 7 and 13 rubrics, with an average of 10.3 per report.

4 Retrieve-Reason-Verify Framework

In this section, we propose a two-stage framework, $R^3V\text{-}EDM$, designed to address the challenges inherent in emergency decision-making, a domain that is inherently safety-critical, where inaccurate or non-compliant outputs can lead to severe consequences, as illustrated in Figure 4.

Real-world decisions must adhere to event-specific conditions and regulations that are often extensive, heterogeneous, and exceed the effective context capacity of LLMs. Additionally, the complexity of the task, and stringent requirements for knowledge dependency, reasoning accuracy, safety, and interpretability make reliable reasoning especially challenging. Motivated by these challenges, this framework retrieves relevant rules from EDM-R^2 and explicitly grounds, verifies, and refines the model's reasoning within these rule constraints, thereby enhancing the accuracy and trustworthiness of emergency decision-making.

4.1 Rule Retrieval and Reasoning

Given an input sample x containing the event information E, task description I, and the optional candidate response set O, our goal is to retrieve the critical rules set R supporting the current decision from the rule base EDM-R^2, and generate a compliant decision answer aligned with the instruction.

We implement a hierarchical, multi-strategy rule retrieval mechanism that incrementally filters and expands rules to form a highly relevant rule set R. The retrieval strategies include: 1) Exact Matching Based on Section: Match when the rule's section field appears in I or O. 2) Text Matching on Trigger Conditions and Actions: Match when the trigger condition or actions fields of a rule appear in I or O. 3) Semantic Similarity-Based Soft Matching: Match when the semantic similarity between trigger conditions or actions and I or O exceeds a predefined threshold. To handle complex multi-step and multi-hop reasoning typical in complex emergency scenarios, we support multi-layer rule retrieval: when a retrieved rule explicitly or implicitly references others, the framework supplements the retrieval of related rules based on the above strategies, progressively expanding the rule set R to cover the entire reasoning chain. After retrieval, rules are ranked by retrieval priority and similarity scores. The input x along with the retrieved rule set R are given to the LLM to generate an initial response decision y, including the decision answer A corresponding to instruction I and the associated reasoning trace T strictly adhering to the retrieved rules while reasoning with event information. We explicitly prompt the LLM that top ranked rules are more relevant.

4.2 Reasoning Verification and Refinement

Although rule-based retrieval significantly strengthens the evidential basis of generated decisions, in high-risk emergency management, single-pass reasoning may still suffer from incomplete chains, insufficient rule usage, or imprecise expression. To further improve reliability and safety, we introduce a reasoning verification module.

During verification, the input x and the initial output y are jointly input to the LLM, which is required to rigorously audit the generation based on event description E, instruction I, candidate set O, and rule set R. The verification focuses on:

1. Whether the reasoning trace is complete, coherent, and covers all key decision points;

2. Whether the final decision strictly complies with and correctly applies the retrieved rules, avoiding violations of regulations or plans.

The model check the initial output and finally

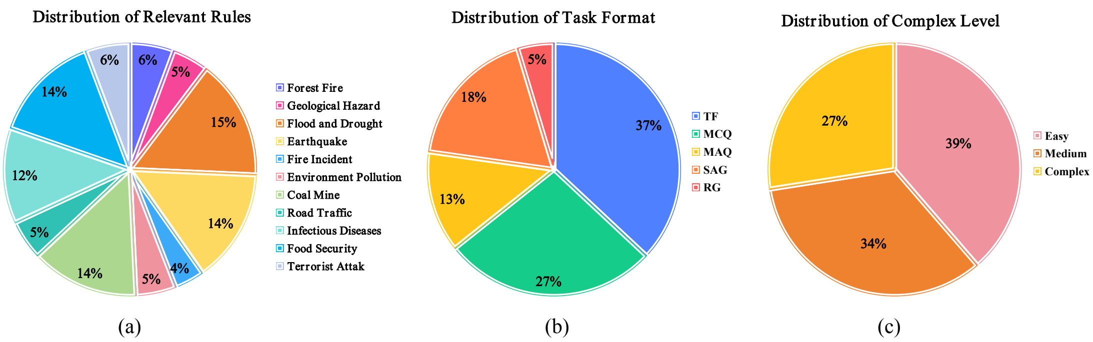

Figure 3: Analysis of EDM-R^2 and EDM-Bench

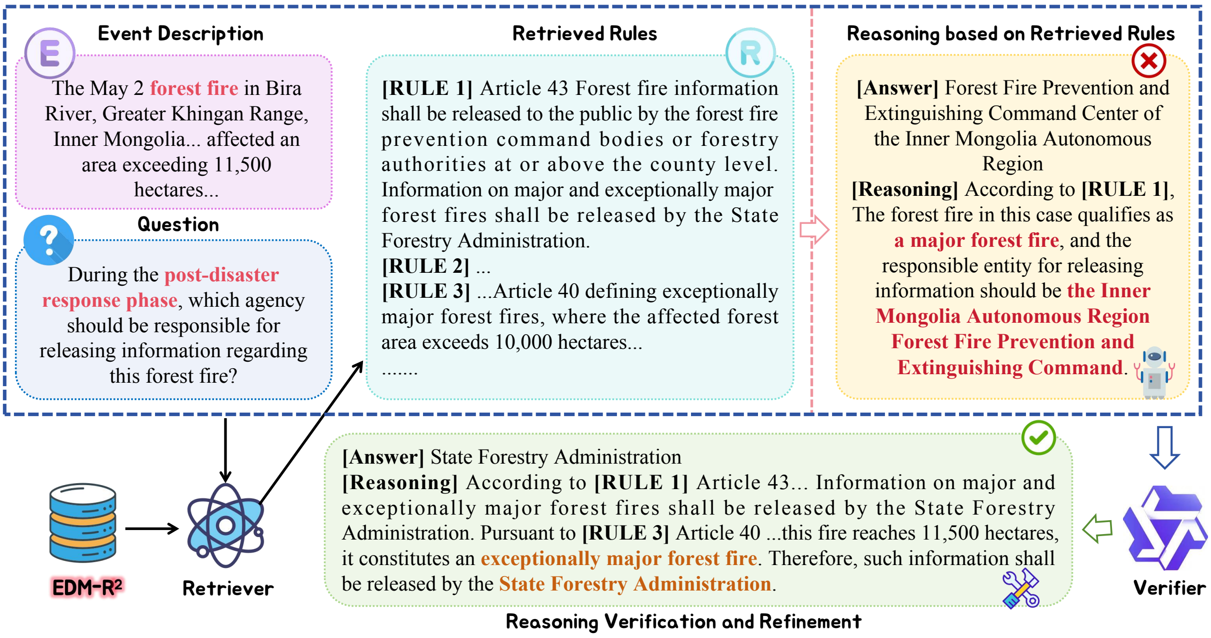

Figure 4: Illustration of $R^3V\text{-}EDM$. An example for SAG task.

produce a verified and refined decision y', comprising the final decision answer A' and rule-based reasoning trace T'. This verification mechanism effectively reduces reasoning biases in complex emergency decisions and enhances compliance, stability, and trustworthiness of the generated results.

5 Experiments

5.1 Models and Evaluation Metrics

Models We selected several mainstream LLMs including open-source and commercial to conduct evaluation on EDM-Bench. We used LLMs to perform decision-making by providing instructions, event information, questions, and answer options as input, without supplying any explicit rules, thereby requiring the models to rely solely on their internal knowledge. This setting is used for comparison with our proposed method. The rules retrieval model was the *paraphrase-multilingual-MiniLM-L12-v2* (Reimers and Gurevych, 2019). Verification model of $R^3V\text{-}EDM$ and evaluation models for SAG and RG are *Qwen2.5-72b-Instruct*.[3]

Evaluation Metrics We employed specific evaluation metrics tailored to different task formats.

TF and MCQ: Use *accuracy* as the evaluation metric, which measures the proportion of correctly predicted answers among all questions.

MAQ: To better reflect the varying risks of different error types in real emergency decisions, we adopt a scoring scheme inspired by manual grading of MAQ exams: each question is worth 4 points; partial omissions score 2 points; any incorrect selections result in zero points. The *average score*, serves as the evaluation metric for MAQ.

SAG: For natural language outputs, simple string matching is insufficient to assess quality and cor-

[3]Experimental details are in Appendix E.

rectness. We employed powerful LLMs directly evaluate semantic consistency between generated content and human annotations, using *agreement rate* as the accuracy metric for this output type.

RG: To evaluate the quality and reliability of generated reports, we annotated detailed rubrics for each report based on expert-defined template modules. The rubrics consist of targeted questions covering key rules, decision authorities, event severity levels, and response actions, each assessing whether a specific criterion is satisfied. The final score is computed by averaging across all reports the proportion of satisfied rubrics per report.[4]

5.2 Main Results

Main results are illustrated in Table 1, highlighting three key findings: **1) Direct use of LLMs is insufficient for reliable emergency decision-making.** Across all evaluated models, performance reveals a significant gap between discriminative accuracy and decision reliability. While models perform competitively on TF and MCQ, their effectiveness sharply declines on more complex, decision-oriented tasks such as MAQ and RG. Notably, RG scores remain below 30% even for leading models like *Claude-Sonnet-4.5*, indicating that general-purpose reasoning alone falls short in producing compliant decisions in safety-critical emergency scenarios. **2) Rule-enhanced reasoning substantially improves decision reliability.** Applying $R^3V\text{-}EDM$ leads to significant performance gains across all task formats, with the largest improvements observed in RG. These results demonstrate that integrating structured regulatory knowledge with constrained inference is essential for converting latent model knowledge into rule-compliant decisions. Importantly, these gains hold across both open-source and proprietary models, indicating the model-agnostic nature of $R^3V\text{-}EDM$. **3) Improvements generalize across task formats and model scales.** Beyond rule-guided generation, $R^3V\text{-}EDM$ improves performance on TF, MCQ, and SAG tasks, demonstrating that constrained reasoning produces more principled outputs without compromising overall proficiency. Furthermore, smaller open-source models enhanced with $R^3V\text{-}EDM$ often match or exceed the performance of larger models used without it. For instance, *Qwen2.5-7b-Instruct* with $R^3V\text{-}EDM$ outperforms the direct use of *Llama-3.1-70b-Instruct*, underscoring the scalability and domain adaptability of our approach.

6 Further Analysis

To gain deeper insights into the dataset and task challenges as well as the effectiveness of our framework, we conducted ablation studies and randomly sampled 95 QA instances with 5 reports for detailed analysis, focusing on common error types, reasoning and answers discrepancies, and model bias and sensitivity issues.[5]

6.1 Ablation Studies for $R^3V\text{-}EDM$

We evaluate the individual contributions of components in $R^3V\text{-}EDM$ across three *Qwen2.5* model scales. Table 2 summarizes the results, showing that both EDM-R^2 and the verifier are essential for optimal performance, though their impact correlates with model size.

Impact of EDM-R^2. Replacing structured rules with vanilla text chunks (-EDM-R^2) leads to a significant performance decline. For *Qwen2.5-7b*, accuracy on TF and MCQ drops by 26.0% and 40.2%, respectively. This trend persists across scales, confirming that explicit rule-based constraints provide critical task-specific guidance that internal LLM knowledge cannot fully substitute.

Effectiveness of Verifier. Removing the verification module (-verifier) consistently degrades results across all metrics. In the 7b model, TF scores decrease from 69.0 to 60.2. While the performance gap narrows in the 72b variant, $R^3V\text{-}EDM$ still outperforms its non-verified counterpart, reinforcing the verifier's role as a robust quality control layer for complex reasoning.

Cross-Scale Robustness. Larger models exhibit greater resilience to the absence of specific components; for example, the 72b model shows a marginal decline in MCQ without rules compared to the catastrophic drop in the 7b model. Nevertheless, the peak performance achieved by the full $R^3V\text{-}EDM$ across all scales demonstrates the synergy between rule-based grounding and iterative verification, particularly in challenging tasks like RG.

6.2 Error Patterns and Analysis

Based on manual inspection of the sampled instances, we categorize errors into three main types: 1) ***Hallucination errors*** *(51%)*, where the model

[4] Details of rubrics and reliability are in Appendix F.

[5] More analyses and cases are in Appendix G and H.

Setting	Model	TF	MCQ	MAQ	SAG	RG
/	glm-4-9b-chat (GLM et al., 2024)	47.8	74.8	1.4	19.9	32.5
	Qwen2.5-7b-Instruct (Qwen et al., 2025)	63.1	73.4	1.8	25.5	37.9
	Llama-3.1-8b-Instruct (Ollama, 2024)	56.7	56.8	0.8	13.6	26.0
	Qwen2.5-32b-Instruct (Qwen et al., 2025)	67.4	81.7	2.5	32.3	37.4
	Qwen2.5-72b-Instruct (Qwen et al., 2025)	57.3	80.2	2.4	29.2	43.1
	Llama-3.1-70b-Instruct (Ollama, 2024)	58.9	73.0	2.0	22.5	27.0
	Llama-3.3-70b-Instruct (Ollama, 2025)	64.3	77.7	2.0	24.0	30.3
	GPT-5.2 (OpenAI, 2025)	60.4	73.9	1.9	35.3	21.5
	Claude-Sonnet-4.5 (Anthropic, 2025)	84.2	84.0	2.7	44.2	18.9
	Deepseek-V3 (Liu et al., 2024)	73.8	90.7	3.4	31.9	29.5
$R^3V\text{-}EDM$	glm-4-9b-chat (GLM et al., 2024)	65.6	83.5	2.4	27.2	30.7
	Qwen2.5-7b-Instruct (Qwen et al., 2025)	69.0	85.3	2.3	33.9	38.0
	Llama-3.1-8b-Instruct (Ollama, 2024)	69.6	84.5	2.1	24.1	32.4
	Qwen2.5-32b-Instruct (Qwen et al., 2025)	74.7	86.0	2.4	38.0	34.7
	Qwen2.5-72b-Instruct (Qwen et al., 2025)	68.1	85.3	2.5	40.1	44.6
	Llama-3.1-70b-Instruct (Ollama, 2024)	74.1	84.4	2.1	32.3	32.6
	Llama-3.3-70b-Instruct (Ollama, 2025)	72.1	87.1	2.4	27.6	34.9
	GPT-5.2 (OpenAI, 2025)	77.0	84.2	2.6	38.5	43.5
	Claude-Sonnet-4.5 (Anthropic, 2025)	85.3	86.4	2.7	44.8	43.1
	Deepseek-V3 (Liu et al., 2024)	70.3	84.2	2.3	35.9	40.0

Table 1: Experimental results cover various settings and task types. In the "/" setting, LLMs perform decision-making using instructions, event information, questions, and answer options without explicit rules, relying solely on internal knowledge. Accuracy is reported for TF and MCQ; Average score (out of 4) for MAQ; Proportion of model outputs consistent with human annotations (as judged by LLM) for SAG; Average rubrics pass rate for RG.

Setting	Model	Setting	TF	MCQ	MAQ	SAG	RG
$R^3V\text{-}EDM$	Qwen2.5-7b-Instruct	$R^3V\text{-}EDM$	69.0	85.3	2.3	33.9	37.9
		-EDM-R^2	43.0	45.1	2.2	16.1	30.7
		-verifier	60.2	77.7	1.8	31.8	36.6
	Qwen2.5-32b-Instruct	$R^3V\text{-}EDM$	74.7	86.0	2.4	38.0	34.7
		-EDM-R^2	36.3	46.6	2.2	39.4	30.7
		-verifier	72.5	83.8	2.4	37.5	34.1
	Qwen2.5-72b-Instruct	$R^3V\text{-}EDM$	68.1	85.3	2.5	40.1	44.6
		-EDM-R^2	65.6	83.8	2.0	30.9	30.9
		-verifier	62.4	84.2	2.5	39.6	43.9

Table 2: Ablation studies for $R^3V\text{-}EDM$.

fabricates non-existent rules during reasoning; 2) ***Fact understanding errors*** *(23%)*, where the rules are correct but misaligned with event-specific information; 3) ***Inappropriate or incomplete rules*** *(26%)*, where the retrieved rules do exist but are not applicable to the given question, or the relevant rules are incomplete and thus insufficient to support a full reasoning and decision-making process.

Incorporating rule retrieval and explicit constraints significantly reduces hallucination errors (23%), with verification and refinement stages further lowering other two error rates (11%).

6.3 Reasoning and Answers Discrepancy

High accuracy alone does not adequately reflect a model's true ability to make correct decisions. Our analysis shows that approximately 73% of correct predictions arise from lucky guessing, where models produce correct answers despite flawed or inconsistent reasoning, particularly in TF and MCQ tasks. This indicates that accuracy-based evaluation substantially overestimates LLM reliability in emergency decision-making, aligning with prior findings (Korbak et al., 2025). By enforcing an evidence-chain mechanism, $R^3V\text{-}EDM$ reduces hallucinated reasoning by 20% and produces over 32% of outputs with explicit, rule-compliant explanations, thereby significantly enhancing interpretability and trustworthiness.

6.4 Bias and Sensitivity Issues

Models exhibit systematic confirmation bias. For example, in TF tasks, although only 27.8% of labels are "correct", *Qwen2.5-72b-Instruct* predicts "Correct" in 85% of cases. When task details are slightly altered, the model often outputs the same "Correct" answer, demonstrating poor sensitivity to critical nuances. Moreover, model performance can vary notably when the same questions and information are presented in different formats, revealing inconsistency in comprehension. Similarly, in MAQ tasks, models tend to prioritize recall over precision, leading to redundant or contradictory selections. These patterns highlight challenges in achieving both sensitivity to detail and consistency across diverse task formulations.

7 Related Work

In this section, we further distinguish our emergency decision-making task from the two most similar tasks: long-document question answering and other high-risk decision-making scenarios.

Long-document QA. Long-document question answering (Bai et al., 2025; Li et al., 2024) primarily evaluates a model's ability to process long contexts and retrieve or synthesize answers grounded in explicitly provided textual evidence. While it may involve multi-hop inference, reasoning is largely text-bound and correctness depends on factual consistency. In contrast, emergency decision-making demands normative and procedural reasoning that integrates event-specific conditions with external regulations and emergency plans, which are often implicit or incomplete in the input. Models must assess rule applicability, resolve regulatory conflicts, infer response levels, and translate policies into compliant actions. Thus, the task focuses on decision validity, regulatory compliance, and risk sensitivity, making it fundamentally more complex and safety-critical than long-document QA.

High-Risk Decision Making. Prior NLP research on high-risk decision-making emphasizing safety and explainability mainly targets at financial (Bartáková et al., 2025; Zhang et al., 2025), healthcare (Sandmann et al., 2025; Li et al., 2025) and legal (Posner and Saran, 2025; Pereira et al., 2025) domains. Work related to emergency decision-making remains limited due to the lack of suitable data and is often confined to specific event types (Chen et al., 2024). Our work addresses this gap by providing a unified task definition, a realistic benchmark, and a rule-augmented reasoning framework for systematic evaluation.

7.1 Iterative Reasoning and Self-Correction

Our framework builds upon the evolving paradigm of self-directed refinement in LLMs. Self-Refine (Madaan et al., 2023) established the effectiveness of iterative feedback loops for improving initial outputs. Similarly, Chain-of-Verification (CoVe) (Dhuliawala et al., 2024) introduced a multi-step process of baseline generation and fact-checking to reduce hallucinations. While these methods focus on general factual or creative improvements, $R^3V\text{-}EDM$ adapts the philosophy of Constitutional AI (Bai et al., 2022), which uses a set of principles to critique and align model behavior, to the domain of emergency response. Unlike generic safety constitutions, our framework employs a dynamic, retrieved set of legal and operational regulations from EDM-R^2 to constrain the model's reasoning within the strict bounds of emergency management protocols.

8 Conclusions

In this paper, we introduce EDM-Bench, the first real-world emergency decision-making dataset with diverse task formats, and EDM-R^2, a structured repository of contingency plans. To address the domain's complexity and safety requirements, we propose $R^3V\text{-}EDM$, a rule-enhanced framework that integrates regulatory knowledge to improve decision reliability and interpretability. Experiments highlight the challenges LLMs face in emergencies and validate our method's effectiveness. Future work will focus on scaling the dataset, refining reasoning, and developing real-time systems for practical deployment.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62476150), and a grant from the Institute for Guo Qiang, Tsinghua University (2019GQB0003). I also would like to thank my advisor for their guidance and our team members for their help.

Limitations

This study primarily aims to explore whether LLMs can reliably support emergency decision-making. Although the prospects are promising, further improvements are necessary to better align with real-world requirements. The main limitations are as follows: **1) Rule coverage, adaptability, and quality:** The current rule repository mainly covers event types present in the dataset, requiring ongoing expansion and maintenance as new event types emerge. Moreover, inconsistencies or ambiguities in regulatory documents can undermine the accuracy and stability of the rule-based retrieval and reasoning framework. **2) Data scale and representativeness:** Despite being constructed from real events, the dataset remains limited in scale and diversity, constrained by historical event availability. The reliance on manually designed question-answer templates and report generation may also limit information completeness. **3) Evaluation of reasoning validity:** Observed "lucky guessing" phenomena and existing manual annotations highlight the need for effective automated methods to assess the completeness and reliability of the model's reasoning paths, which will be a focus of future work.

Ethical Considerations

In this section, we discuss the ethical considerations of this work as follows: (1) **Data Sources and Privacy.** The EDM-Bench dataset is built from publicly available regulatory documents and official emergency reports. No personal or sensitive information is included, ensuring compliance with privacy standards. (2) **Intended Use and Risks.** EDM-Bench and $R^3V\text{-}EDM$ are intended to support research on reliable and interpretable emergency decision-making systems. Given the safety-critical nature of the domain, model outputs should be used with expert oversight to avoid potential risks from erroneous decisions. (3) **Fair Evaluation.** To promote fair and consistent benchmarking, data and evaluation codes will be openly shared. (4) **Annotator Welfare.** Data annotation and verification were carried out by trained annotators under fair and ethical working conditions. Specifically, we recruited eight graduate students with backgrounds in NLP to perform the annotation tasks. (5) **AI Assistance.** We used AI Assistants (ChatGPT) for language refinement for this paper.

References

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664.

Gabriela Pajtinková Bartáková, Ahmad Almadhor, Amna Qayyum, Kainaat Abeer, Aman Durrani, and 1 others. 2025. Evaluating the capacity and limitations of generative ai in financial decision making. *Computer Standards & Interfaces*, 93:103965.

Minze Chen, Zhenxiang Tao, Weitong Tang, Tingxin Qin, Rui Yang, and Chunli Zhu. 2024. Enhancing emergency decision-making with knowledge graphs and large language models. *International Journal of Disaster Risk Reduction*, 113:104804.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. Crisisltlsum: A benchmark for local crisis event timeline extraction and summarization. *arXiv preprint arXiv:2210.14190*.

Demian Gholipour Ghalandari, Chris Hokamp, John Glover, Georgiana Ifrim, and 1 others. 2020. A large-scale multi-document summarization dataset from the wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, and 37 others. 2024. Chatglm:

A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Yun He, Wenzhe Li, Hejia Zhang, Songlin Li, Karishma Mandyam, Sopan Khosla, Yuanhao Xiong, Nanshu Wang, Xiaoliang Peng, Beibin Li, and 1 others. 2025. Advancedif: Rubric-based benchmarking and reinforcement learning for advancing llm instruction following. *arXiv preprint arXiv:2511.10507*.

Zhuowei Hu, Huili Gong, and Liying Zhu. 2007. Fast flooding information extraction in emergency response of flood disaster. In *Proceedings of the ISPRS workshop on updating geo-spatial databases with imagery and the 5th ISPRS workshop on DMGISs, Urumchi*, pages 28–29.

Ahrii Kim. 2025. Rubric-mqm: Span-level llm-as-judge in machine translation for high-end models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 147–165.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, and 1 others. 2025. Chain of thought monitorability: A new and fragile opportunity for ai safety. *arXiv preprint arXiv:2507.11473*.

Jia Li, Zichun Zhou, Han Lyu, and Zhenchang Wang. 2025. Large language models-powered clinical decision support: enhancing or replacing human expertise?

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. Loogle: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jiaheng Liu, Dawei Zhu, Zhiqi Bai, Yancheng He, Huanxuan Liao, Haoran Que, Zekun Wang, Chenchen Zhang, Ge Zhang, Jiebin Zhang, and 1 others. 2025. A comprehensive survey on long context language modeling. *arXiv preprint arXiv:2503.17407*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594.

Heng Mu, Peng Wu, and Wenyi Su. 2024. Construction of knowledge graph for emergency resources. *International Journal of Intelligent Systems*, 2024(1):6668559.

Ollama. 2024. Llama 3.1. https://ollama.com/library/llama3.1.

Ollama. 2025. Llama 3.3. https://ollama.com/library/llama3.3.

OpenAI. 2025. Update to gpt-5 system card: Gpt-5.2.

Aditya Pathak, Rachit Gandhi, Vaibhav Uttam, Yashwanth Nakka, Aaryan Raj Jindal, Pratyush Ghosh, Arnav Ramamoorthy, Shreyash Verma, Aditya Mittal, Aashna Ased, and 1 others. 2025. Rubric is all you need: Enhancing llm-based code evaluation with question-specific rubrics. *arXiv preprint arXiv:2503.23989*.

Jayr Pereira, Andre Assumpcao, Julio Trecenti, Luiz Airosa, Caio Lente, Jhonatan Cléto, Guilherme Dobins, Rodrigo Nogueira, Luis Mitchell, and Roberto Lotufo. 2025. Inacia: Integrating large language models in brazilian audit courts: Opportunities and challenges. *Digital Government: Research and Practice*, 6(1):1–20.

Eric A Posner and Shivam Saran. 2025. Judge ai: Assessing large language models in judicial decision-making. *University of Chicago Coase-Sandor Institute for Law & Economics Research Paper*, (2503).

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sarah Sandmann, Stefan Hegselmann, Michael Fujarski, Lucas Bickmann, Benjamin Wild, Roland Eils, and Julian Varghese. 2025. Benchmark evaluation of deepseek large language models in clinical decision-making. *Nature Medicine*, pages 1–1.

Wenxin Su, Linyan Chen, and Xin Gao. 2022. Emergency decision making: A literature review and future directions. *Sustainability*, 14(17):10925.

Jianzhuo Yan, Lihong Chen, Yongchuan Yu, Hongxia Xu, Qingcai Gao, Kunpeng Cao, and Jianhui Chen. 2022. Emergeventmine: End-to-end chinese emergency event extraction using a deep adversarial network. *ISPRS International Journal of Geo-Information*, 11(6):345.

Qingyuan Zhang, Yuxi Wang, Cancan Hua, Yulin Huang, and Ning Lyu. 2025. Knowledge-augmented large language model agents for explainable financial decision-making. *arXiv preprint arXiv:2512.09440*.

Zi-Xin Zhang, Liang Wang, and Ying-Ming Wang. 2018. An emergency decision making method based on prospect theory for different emergency situations. *International Journal of Disaster Risk Science*, 9(3):407–420.

Mengna Zhu, Kaisheng Zeng, Mao Wang, Kaiming Xiao, Lei Hou, Hongbin Huang, and Juanzi Li. 2025. Eventsum: A large-scale event-centric summarization dataset for chinese multi-news documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26138–26147.

A Basic Concepts

Emergency According to the definitions provided in relevant Chinese laws and regulations, the emergencies involved in this paper refer to sudden occurrences that have caused or may cause serious social harm and require emergency response measures, including natural disasters, accident incidents, public health emergencies, and social security incidents, such as earthquakes, geological hazards, and terrorist attacks.[6] Such events are characterized by suddenness, harm, and urgency, posing challenges for monitoring, rapid response, and decision-making.

Emergency Decision Making Emergency decision making refers to the process where government agencies or emergency authorities, guided by laws, plans, and expertise, assess an incident's development and make quick, reasonable, and actionable response decisions. This process involves severity assessment, reporting, response activation, resource coordination, command issuance, and investigation. This study focuses on understanding events from natural language and reasoning about decisions consistent with real-world emergency handling.

B Taxonomy

Based on relevant emergency regulations and the corpus collection, we formed a taxonomy which focused on four categories (natural disasters, accident incidents, public health emergencies, and social security incidents) encompassing 11 major common event types, covering a wide range of emergency scenarios and are clearly defined in official response plans. The taxonomy used for data construction is shown in Table 3.

C Details of EDM-R^2 Construction

Construction method can be seen in Figure 5.

Data Quality Control and Analysis To ensure the quality of the rule base, we conducted mannually spot checks on the rules, focusing on the correspondence among trigger conditions, decision authorities, and response actions. The data pass rate reached 98%, and identified issues were promptly corrected during the inspection.

[6]https://www.mem.gov.cn/fw/yjya/

D Details of EDM-Bench Construction

D.1 Construction Details of EDM-QA

The prompt used for candidate TF questions generation can be seen in Figure 6.

The prompt used to convert TF questions into MCQ questions can be seen in Figure 7. Other prompts are similar.

Example data can be seen in Figure 8.

D.2 Construction Details of EDM-Report

D.2.1 Report Template

The report data template was developed by experts based on real response reports, emergency plans, regulations, and other relevant materials. Example template for forest fire structured report generation consists of following modules:

Description of the Incident: Including the time, location, cause, main developments, and resulting impacts of the event.

Incident Severity and Risk Assessment: Basis for evaluation and assessment results.

Reporting and Notification of Key Information: Relevant departments involved in reporting and notification.

Determination of Emergency Response Level: Supporting rationale and the assigned response level.

Primary Emergency Response Measures and Action Plans: Command structures for firefighting, related response agencies, and specific actions taken.

Follow-up Actions: Post-incident management after the emergency is under control, including fire assessment, investigation and accountability for the fire, interviews and rectification measures, responsibility assignment, work summary, and recognition or awards.

D.2.2 Report Data Example

According to the above template, a report data example is as follows.

 # id: 0001
 # category: Natural Disaster
 # event type: Forest and Grassland Fire
 # type: Report Generation Task
 # event: May 9 Yajiang Mountain Fire

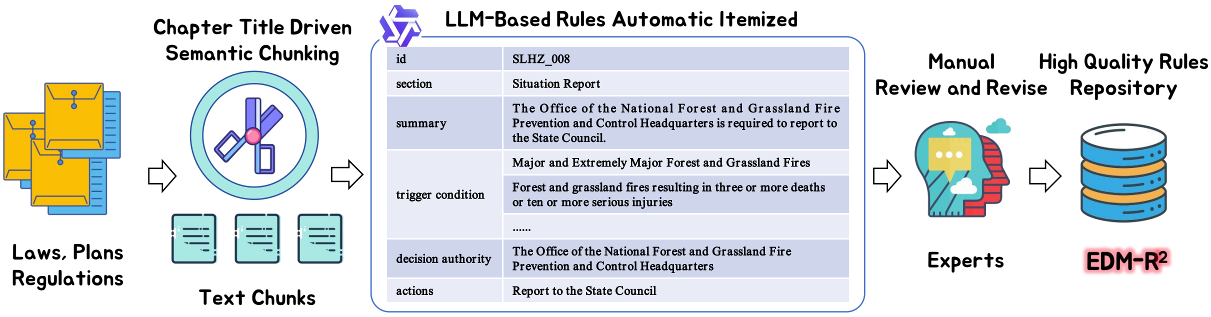

Figure 5: Overview of the Construction Method. Collecting, Segmenting, Parsing, and Structuring Legal Documents into the Emergency Rule Repository EDM-R^2.

<instruction>
You are an expert proficient in natural language processing and emergency response tasks, skilled at integrating regulations to make emergency decisions for sudden events, with rich background knowledge. Please help me design some true/false questions based on the provided events and regulations. The questions should correspond to different difficulty levels.
My difficulty design requirements include three types:

Easy
Can be directly inferred from a specific regulation clause, such as the "three no's" in disaster response (direct retrieval and matching).

Medium
Requires simple reasoning by combining event information, performing some calculations, or understanding the event details to determine which regulation clause to apply.

Complex
Requires integrating multiple regulation clauses, involving multi-hop reasoning.
Facts are mixed: most information in the question stem is correct, but parts contain misleading or distracting information (such as replaced key agencies, quantities, or figures).
Compound questions.

Please help me design 50 candidate true/false question-answer pairs, and return the reasoning process that integrates the relevant regulations.

Requirements:
1. Avoid generating questions that obviously violate common sense, i.e., try not to create questions answerable solely based on common sense without any regulations or background knowledge.
2. Mention the chapter or section titles of the regulations in the question stems to facilitate later retrieval.
3. Design the questions and answers by combining event information and regulation rules.
4. The reasoning process must include the complete regulatory information used.

<input>
[event description] Text describing the basic situation of the event
[rules] Rules randomly retrieved from our constructed rule repository EDM-R^2.

<output>
[type] task format, TF question
[event description] Text describing the basic situation of the event
[question] The question related to emergency response generated based on regulations
[options] A. correct B. Incorrect
[answer] A / B
[reasoning] Reasoning process based on the given rules

Figure 6: Relevant prompt for candidate TF questions generation.

<instruction>
You are an expert in natural language processing and emergency response tasks, with strong expertise in applying regulations to make emergency decisions for sudden incidents and rich background knowledge. Please help me rewrite the true/false questions into multi-choice questions.

The construction of incorrect options can be based on the following methods:

- Modifying key information such as responsible agencies or response levels
- Omitting critical information
- Confusing response measures applicable to different scenarios
- Mixing correct and incorrect statements within the same option

Requirements:
1. Avoid generating options that obviously violate common sense; that is, try to avoid questions that can be answered based on common sense alone without any regulatory or domain knowledge.
2. Mention the chapter or section titles of the regulations in the question stems to facilitate subsequent retrieval.
3. Design the questions and answers by integrating both event information and regulatory rules.
4. Ensure that the options are as distractive as possible.
5. The regulatory information used in the reasoning process must be complete.

<input>
TF data

<output>
converted candidate MCQ data

Figure 7: Relevant prompt for converting TF questions into candidate MCQ questions generation.

 # time: May 9, 2024

 # location: Yajiang County, Ganzi Prefecture, Sichuan Province

 # event description: At 17:30 on May 9, 2024, a forest fire broke out in Yajiang County, Ganzi Prefecture, Sichuan Province. Due to sustained strong winds and complex terrain, the fire rapidly spread to surrounding forest areas, with large amounts of smoke entering urban areas causing significant local air quality deterioration. The fire posed a temporary threat to urban safety, involving a liquefied gas storage station (with approximately 180 tons in reserve), three refueling facilities, two primary and secondary schools, and a large logistics warehouse.

By the evening of May 12, the burned area reached approximately 1,500 hectares, with about 260 hectares of damaged forest. The fire continued to show signs of spreading. The State Council attached great importance and issued important directives.

By around 11:00 on May 14, the main fire and northern line flames were basically extinguished, and the situation was preliminarily controlled.

By about 10:00 on May 15, all visible flames on site were extinguished, and the response shifted to cleaning residual fires, hazard investigation, and monitoring for rekindling. Emergency rescue forces gradually withdrew by May 18. Post-investigation determined the fire was caused by a power transmission line fault under hot and dry weather conditions, resulting in direct economic losses of approximately 72 million RMB. No casualties were reported.

 # question: Please generate an emergency response report covering the following aspects: *description of the incident*, *incident severity and risk assessment*, *reporting and notification of key in-*

<Event Description>
..The May 2 forest fire in Greater Khingan Range, Inner Mongolia... affected an area exceeding 11,500 hectares, threatening the lives and property of over 30,000 residents...

<Question>
During the post-disaster response phase, information regarding this forest fire should be release by Forestry Department of the Inner Mongolia Autonomous Region.
<Options>
A. Correct
B. Incorrect
<Answer>
B
<Reasoning>
According to Article 43 of the Forest Fire Prevention Regulations...Information on major and exceptionally major forest fires shall be released by the State Forestry Administration. Pursuant to Article 40 defining exceptionally major forest fires, where the affected forest area exceeds 10,000 hectares and this fire reaches 11,500 hectares, it constitutes an exceptionally major forest fire. Therefore, such information shall be released by the State Forestry Administration. Decision is incorrect!

<Question>
During the post-disaster response phase, which agency should be responsible for releasing information regarding this forest fire?
<Options>
A. Forest Fire Prevention and Extinguishing Command Center of the Inner Mongolia Autonomous Region
B. Forestry Department of the Inner Mongolia Autonomous Region
C. People's Government of Hulunbuir City
D. State Forestry Administration
<Answer>
D
<Reasoning>
Similar to TF data.

<Question>
During the risk assessment phase, what level should be assigned to this fire? In the post-disaster response phase, which agency is responsible for releasing information about this forest fire?
<Options>
A. Major Forest Fire
B. Exceptionally Major Forest Fire
C. Forestry Department of the Inner Mongolia Autonomous Region
D. People's Government of Hulunbuir City
E. State Forestry Administration
<Answer>
BE
<Reasoning>
Similar to TF data.

<Question>
During the post-disaster response phase, which agency should be responsible for releasing information regarding this forest fire?
<Answer>
State Forestry Administration
<Reasoning>
Similar to TF data.

Figure 8: Example data of EDM-QA. Blue: TF; Yellow: MCQ; Orange: MAQ; Green: SAG.

formation, *determination of emergency response level*, *main emergency response measures and action plans*, and *follow-up actions*.

 # modules:

- Description of the incident: time, location, cause, main developments, and impact
- Incident severity and risk assessment: legal basis and derived severity conclusion
- Reporting and notification of key information: relevant legal basis and reporting departments
- Determination of emergency response level: rationale and assigned response level conclusion
- Main emergency response measures and action plans: related response agencies and action arrangements
- Follow-up actions: subsequent measures including fire assessment, investigation and accountability, and work summary

 # answer:

1. Description of the Incident (time, location, cause, main developments, results, and impact)
At 17:30 on May 9, 2024, a forest fire occurred in Yajiang County, Ganzi Prefecture, Sichuan Province. The fire was triggered by a power transmission line fault under hot and dry weather conditions. By the evening of May 12, the fire was still spreading, prompting high-level attention and directives from the State Council. By approximately 11:00 on May 13, the main fire and northern line flames were largely extinguished, with preliminary control achieved. By about 10:00 on May 14, all visible flames were extinguished, shifting response efforts to residual fire cleanup, hazard investigation, and monitoring. Emergency rescue teams gradually withdrew by May 18. The fire caused forest resource damage and notable environmental and economic impacts, with direct economic losses of about 72 million RMB. The burned area was approximately 1,500 hectares, with about 260 hectares of forest damaged. No casualties were reported.

2. Incident Severity and Risk Assessment (legal references and conclusion)
According to Article 40 of the Forest Fire Prevention and Control Regulations, forest fires are classified as ordinary, larger, major, and especially major based on the affected forest area and casualties:
   - Ordinary forest fire: affected area under 1 hectare or other forest land fires, or 1 to 3 deaths, or 1 to 10 serious injuries;
   - Larger forest fire: affected area 1 to 100 hectares, or 3 to 10 deaths, or 10 to 50 serious injuries;
   - Major forest fire: affected area 100 to 1000 hectares, or 10 to 30 deaths, or 50

to 100 serious injuries;

- Exceptionally major forest fire: affected area over 1000 hectares, or over 30 deaths, or over 100 serious injuries.

This incident affected between 100 and 1000 hectares with no casualties and is therefore classified as a major forest fire.

3. Reporting and Notification of Key Information
According to Section 5.2 of the National Forest and Grassland Fire Emergency Plan, local forest (grassland) fire prevention and control command agencies must report fire information timely, accurately, progressively, and in a standardized manner, under the principle of "report all fires". The National Forest and Grassland Fire Prevention Command Office reports the following to the State Council:

   - Major and exceptionally major forest and grassland fires;
   - Fires causing three or more deaths or ten or more serious injuries;
   - Fires threatening residential areas or critical infrastructure;
   - Fires within 5 kilometers of national borders threatening resources of China or neighboring countries;
   - Other important fires as assessed.

   Article 43 of the Forest Fire Prevention and Control Regulations stipulates fire information must be released by county-level or higher forestry departments or fire prevention command agencies, with major and especially major fires announced by the State Forestry Administration. This incident qualifies as a major fire, thus Yajiang County, Ganzi Prefecture, and Sichuan forestry fire command must report progressively and accurately, with the national command office reporting to the State Council, and information released by the State Forestry Administration.

4. Determination of Emergency Response Level
According to Chapter 6 of the National Forest and Grassland Fire Emergency Plan, four response levels (IV, III, II, I) are set at the national level based on fire severity, fire progression, and local firefighting conditions, with provincial and other authorities implementing corresponding measures. With approximately 1,500 hectares burned and 260 hectares of damaged forest, and no casualties, but with fire spreading for over 72 hours from May 9 to May 12 evening without effective control, a Level II emergency response was initiated.

5. Main Emergency Response Measures and Action Plans
According to Section 3.3 of the National Forest and Grassland Fire Emergency Plan, firefighting is commanded by local forest (grassland) fire prevention agencies. Multiple fires or fires crossing administrative boundaries are commanded by higher-level agencies. Cross-province fires of varying severities are directed by county, city, provincial, or national command accordingly. This fire did not cross provincial boundaries and involved a single fire site, so the Yajiang County Forestry Fire Command directed the response. Under Level II response and Article 37 of the Forest Fire Prevention Regulations, the following measures apply:

   - The National Forest and Grassland Fire Command organizes joint consultations, analyzes fire risk, studies firefighting measures, and dispatches work groups to guide firefighting;
   - Reinforcement by local specialized firefighting teams, national comprehensive firefighting rescue teams, and aerial firefighting support;
   - Coordination and deployment of military and paramilitary forces across regions;
   - Weather forecasting and artificial precipitation by meteorological authorities;
   - Protection of important targets and hazardous sources;
   - Priority transport support for personnel and supplies;
   - Emergency communication guarantees;
   - Evacuation shelters and disaster relief by civil affairs departments;
   - Security maintenance by public security agencies;
   - Material supply, medical rescue, and epidemic prevention by commerce and health departments;
   - Timely press conferences and public information management.

6. Follow-up Actions
According to Article 41 of the Forest Fire Prevention and Control Regulations, forestry departments above the county level shall investigate and assess causes, responsible parties, affected forest area and stock, casualties, and other economic losses, and report to local governments, which shall determine responsible units and individuals and handle accordingly. Loss assessment standards are formulated by the State Forestry Administration with relevant departments. Section 8.5 of the National Forest and Grassland Fire Emergency Plan requires fire prevention command agencies to summarize and analyze causes and lessons, proposing improvements. For fires of high importance or wide social impact, the National Command submits a firefighting work summary to the State Council. Therefore, Yajiang County, Ganzi Prefecture, and Sichuan forestry authorities shall conduct investigation and assessment, submit reports, and assign responsibility for this incident. The State Council has issued directives and will receive a final summary. The fire was caused by power line faults under specific wind conditions. Relevant power facility maintenance units bear supervisory responsibility. Forestry and power management authorities should strengthen facility safety management and fire prevention duties to prevent recurrence.

 # rubrics:

Q1: Does the report include all essential elements, such as a description of the basic situation of the event, event level and risk assessment, reporting and notification of key information, emergency response level determination, main emergency response measures and action arrangements, and follow-up handling module?

Q2: Does the event level and risk assessment refer to Article 40 of the Forest Fire Prevention and Control Regulations?

Q3: Is the event level and risk assessment concluded as a major forest fire?

Q4: Is the information report submitted by the National Forest Fire Prevention and Control Command Office to the State Council?

Q5: Is the forest fire information released to the public by the forestry authority under the State Council?

Q6: Is the emergency response level set to Level II?

Q7: Is the fire-fighting command led by the Forest Fire Prevention Command of Yajiang County, Sichuan Province?

Q8: In the emergency response measures and action arrangements, is a working group organized by the State Council dispatched to the fire site to coordinate and guide fire-fighting efforts?

Q9: In the emergency response measures, does the meteorological department guide and supervise local artificial weather modification operations based on fire site meteorological conditions?

Q10: In the follow-up handling, does the forestry authority investigate and assess the fire cause, casualties, economic losses, and submit a report to the people's government?

Q11: In the follow-up handling, does the people's government determine responsible units and persons based on the investigation report and handle them according to law?

Q12: In the follow-up handling, does the National Forest and Grassland Fire Prevention and Control Command submit a summary report of the fire-fighting work to the State Council?

D.3 Quality Control

For EDM-QA, to ensure annotation quality, we implement a strict quality control mechanism throughout the data construction process, combining LLM-based scoring and manual inspection. The process includes multi-dimensional LLM evaluation, multi-round cross-validation, hierarchical review, and project manager spot checks with feedback.

First, the data are annotated in batches by event category, with each annotator responsible for a fixed event type to maintain consistency in understanding events and regulations. The project manager then applies LLM-based multi-dimensional scoring. Samples scoring below a predefined threshold are treated as low-quality and subjected to detailed review. Common issues are summarized and fed back to annotators, and the annotation guidelines are updated accordingly. Once high-quality samples (above the threshold) exceed 95%, multi-round cross-validation is conducted, where annotators review data from other types to identify factual errors and resolve ambiguities through discussion.

After cross-validation, the manager performs random spot checks and corrections. If the pass rate exceeds 95%, the annotation process is finalized; otherwise, the cross-validation step is repeated.

```
<instruction>
You are a senior review expert. Based on the provided event description, question, options, and relevant regulations, please strictly verify the correctness of the initial decision.
Your tasks are:
• Verify: Determine whether the initial decision (answer and reasoning) is correct, and whether the reasoning process is complete and complies with the regulations.
• Correct: If any errors or insufficient reasoning are found, please make corrections.
• Final output: Regardless of whether corrections were made, please return the final and most accurate decision result.
---------------------------- Information ----------------------------
## Event Description: {event_description}
## Question: {question}
## Options: {options}
## Most Relevant Regulations: {rules_text}
------------------------- Initial Decision -------------------------
## Initial Answer: {initial_answer}
## Initial Reasoning: {initial_reasoning}

Output Requirements:
1. You must strictly output in the following JSON format, including the final answer and reasoning.
{
"answer": "Final answer (e.g., A or ACD or specific text)",
"reasoning": "Detailed verification and correction process, and the complete basis for the final answer"
}
2. The reasoning must clearly state whether corrections were made and the basis for corrections (if any).
3. Your output will be regarded as the final result.
--------------------------------------------------------------------------------
<input>
Event Description | Question | Options (optional) | Relevant Rules
Initial Decision (Answer & Reasoning)
--------------------------------------------------------------------------------
<output>
Final Decision (Answer & Reasoning)
```

Figure 9: Relevant prompt for verification and refinement process of $R^3V\text{-}EDM$.

Through this multi-layered quality control process, the final dataset achieves high accuracy and consistency, and the final pass rate is 97%.

For EDM-Report, quality control is primarily conducted through manual inspection. Annotators from different event types perform cross-checking and correction of the reference reports, followed by a final review by the project manager to ensure data quality. The report data are finalized only after passing manager verification.

D.4 Data Details

EDM-Bench details can be seen in Table 4.

E Experimental Details

The experimental results reported in the paper are based on single-run experiments. All experiments were conducted on NVIDIA A100 GPUs, with a total computational cost of approximately 100 GPU-hours.

The threshold for semantic similarity is *0.6*. Temperature is *0.0*. Top-k rules number is *5*. Max tokens for TF, MCQ, MAQ, SAG are *1024*, for RG are *4096*.

The prompt used for verification and refinement can be seen in Figure 9.

F Details of Rubric-based Evaluation

F.1 Rubrics Annotation

We formulate questions based on the templates summarized by experts and the modules involved in the data, in conjunction with human-written answers. These questions are centered on the modules and their corresponding regulations, with particular emphasis on decision-critical information—such as applicable regulations, responsible authorities, event classification, and prescribed actions—to determine whether a report satisfies these criteria and thereby enable a comprehensive assessment of the accuracy of the reported information. An illustrative example of rubric annotation is provided in Section D.2.2.

F.2 Reliability of Rubric-based Evaluation

To assess the consistency between model-based and human evaluations, we conducted an analysis on a sampled subset of 5 reports with 49 rubrics. For each report, both human annotators and the model evaluated the satisfaction of all rubric items as binary decisions. Across all rubric judgments within this sample, the overall agreement rate between model predictions and human annotations was 95.9%, indicating substantial alignment between automatic and human evaluations.

F.3 Evaluation

We used *Qwen2.5-72b-Instruct* for evaluation. The relevant prompt and evaluation example can be seen in Figure 10.

G More Analysis

To obtain a more comprehensive understanding of the data, tasks, and our method, we conducted additional experiments for further analysis.

G.1 Performances across Complex Levels

In this section, we analyze the model's performance on QA data with varying difficulty levels, as shown in Figure 11.

Our analysis reveals a clear performance degradation as task difficulty increases across all models and settings, confirming that higher complexity poses greater challenges for large language models. Under the LLM directly using setting, the performance gap between easy and complex levels is notably large in several tasks, especially for multi-choice question answering (MCQ) and question answering with explanation (SAG).

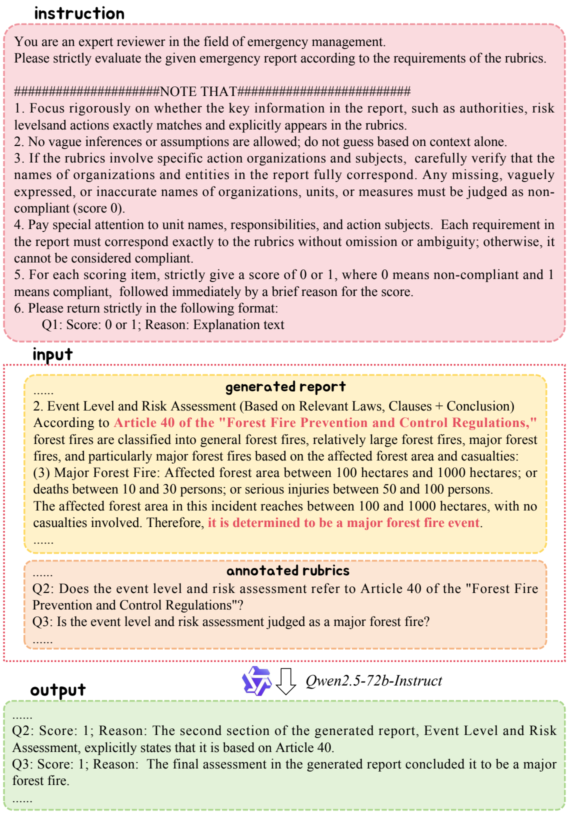

Figure 10: Relevant prompt and rubric-based evaluation example for report evaluation.

When applying our proposed method, the overall model performance improves consistently across difficulty levels. More importantly, the method significantly reduces the performance gap between easy and complex levels for most tasks and models. For example, in MCQ and SAG tasks, the gap narrows substantially, indicating enhanced robustness and adaptability of the models to complex inputs. However, for tasks like TF (task-focused assessment), the gap reduction is less consistent, suggesting that some tasks remain more sensitive to complexity despite the method.

Moreover, larger models (Qwen2.5-32b and Qwen2.5-72b) tend to benefit more from the method in terms of narrowing difficulty-related performance disparities, implying a synergistic effect between model scale and the proposed approach.

In summary, the proposed method effectively mitigates the negative impact of increasing difficulty on model performance, particularly in reasoning-intensive tasks. This suggests the approach improves not only absolute performance but also the stability of model behavior across varying complexity levels.

H Case Study

H.1 Error Patterns

In this section, we present concrete cases for each error pattern to facilitate an in-depth analysis. The main error patterns include hallucination errors, inappropriate or incomplete rules, and factual understanding errors.

Hallucination Errors In the example shown in Figure 12(a), the content of *Article 21* of the Regulations on the Prevention and Control of Geological Disasters does not correspond to the actual regulation, indicating a fabricated or incorrect citation. Furthermore, this article is not applicable to the issue at hand and does not explicitly state a requirement to report within 4 hours.

Fact Understanding Errors As shown in the example in Figure 12(c), there is a misunderstanding of the rule details: the event information states that 68,000 people temporarily switched to backup water sources, not that they were evacuated or relocated. Additionally, the economic loss did not exceed 100 million yuan. Therefore, the criteria for an extraordinarily major environmental incident were not met, and this does not constitute an extraordinarily major environmental incident.

Inappropriate or Incomplete Rules As shown in the example in Figure 12(b), there are specific regulations regarding geological disasters, and the referenced rule is not appropriate. The correct rule and reasoning are as follows:

According to Article 28 of the Regulations on the Prevention and Control of Geological Disasters, units and individuals who discover geological disaster risks or disasters shall immediately report to the local people's government or the land and resources authority.

The key institutions and departments involved in the report are incorrectly identified.

H.2 Reasoning and Answers Discrepancy

Accuracy alone does not fully reflect the model's capability. By comparing the model's reasoning output with human-annotated reasoning, we found a large amount of invalid reasoning that is inaccurate or incomplete. As shown in the Figure 13, the complete human-annotated reasoning path should be: Forest and grassland fire suppression work is commanded by the local forest (grassland) fire prevention and control agency. When three or

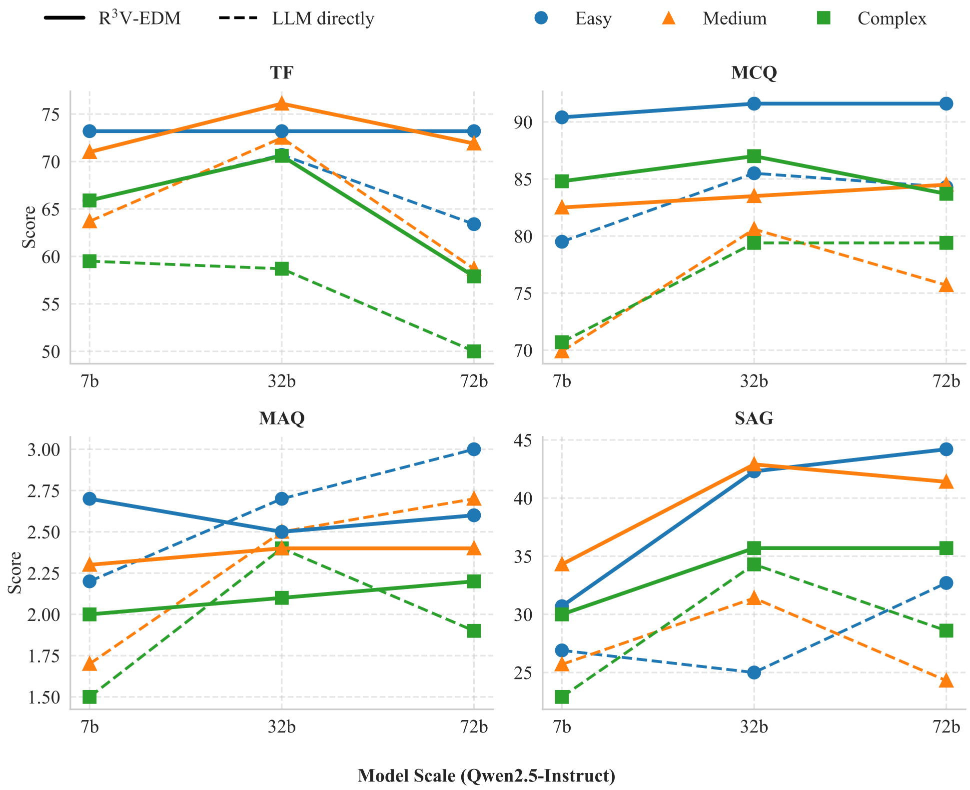

Figure 11: Different Performances across Complex Levels.

more fires occur simultaneously, or when a single fire spans two administrative regions, the command should be taken over by the higher-level forest (grassland) fire prevention and control agency. County-level and above local governments establish forest (grassland) fire prevention and control agencies according to the principle of "correspondence at each level", responsible for organizing, coordinating, and guiding forest and grassland fire prevention and control within their jurisdictions. In this incident, since three fire zones have formed, the command should be led by the higher-level forest and grassland fire prevention and control agency. As such agencies are only established at county level and above, the command here should be by the Garze Prefecture Forest and Grassland Fire Prevention and Control Headquarters. Therefore, the decision is correct. The model, however, did not distinguish the levels of the commanding agencies.

H.3 Bias and Sensitivity Issues

In this section, we provide a concrete example to illustrate that the model tends to always answer "correct" but lacks the ability to discern details for TF task. When key information in the question stem is slightly perturbed, the model fails to recognize it. As shown in Figure 14, regarding the same event, the model considers both large-scale and medium-scale disaster classifications as correct, rendering the reasoning process invalid and of low quality.

As for sensitivity issues, as shown in Figure 15, the decision-related questions all focus on which agency is responsible for delineating the geological hazard danger zone after the occurrence of a geological disaster. However, different question formats use different regulations based on the model's internal knowledge. Among the TF, MCQ, and SAG types, only the Multiple Choice Question were answered correctly, though they were still accompanied by incorrect rules.

The correct human-written reasoning is as follows:

According to Chapter 4 of the National Emergency Plan for Sudden Geological Disasters, titled Classification of Geological Disaster Risk and Disaster Levels, a geological disaster risk is classified as extra-large if the number of people threatened and needing relocation exceeds 1,000, or if the potential economic loss exceeds 100 million RMB. Similarly, a geological disaster is classified as extra-large if there are more than 30 deaths or direct economic losses exceed 10 million RMB.

Given the reported economic loss of 29.37 mil-

⟨Event Description⟩
The December 4 Qiaojia village ground fissure incident is a geological disaster event that occurred on December 4, 2024, in Renhe Village, Jintang Town, Qiaojia County, Zhaotong City, Yunnan Province, China. Villagers discovered ground cracks while preparing a village banquet. The crack width rapidly expanded from a few centimeters in the early morning to 30 centimeters by night, with localized subsidence reaching 40 centimeters. Thanks to timely discovery, all 79 households and 174 people in the village were relocated to hotels that night, resulting in no casualties, although livestock were not evacuated in time....

⟨Question⟩
After receiving the disaster report, the land and resources authority of the Qiaojia County People's Government must promptly report to the county-level people's government and the land and resources authority of the municipal people's government within 4 hours.

⟨Answer⟩
A
⟨Reasoning⟩
According to Article 21 of the Regulations on the Prevention and Control of Geological Disasters, after receiving a report of a geological disaster in the locality, the land and resources authority of the local people's government at or above the county level shall immediately dispatch personnel to the site to conduct an on-site investigation, take effective measures to prevent the disaster from expanding, and report to the higher authorities in accordance with the provisions of the State Council's land and resources authority. Therefore, the land and resources authority of the Qiaojia County People's Government, upon receiving the disaster report, must promptly report to the county-level people's government and the land and resources authority of the municipal people's government within 4 hours. This practice complies with the relevant regulatory requirements.

(a) Halluciniation Errors

⟨Event Description⟩
At approximately 11:50 AM on February 8, 2025, a sudden landslide occurred in Group 2, Jinping Village, Mu'ai Town, Junlian County, Yibin City, Sichuan Province, burying several houses and trapping people...

⟨Question⟩
After the occurrence of this geological disaster, the units and individuals who discover the disaster should immediately report it to the Junlian County People's Government and the emergency management authorities.

⟨Answer⟩
A
⟨Reasoning⟩
According to Article 38 of the Law of the People's Republic of China on Emergency Response to Public Emergencies, units and individuals who discover a sudden incident shall immediately report it to the local people's government or relevant competent authorities. Therefore, after the occurrence of this geological disaster, the units and individuals who discover the disaster should immediately report it to the Junlian County People's Government and the emergency management authorities.

(b) Inappropriate or Incomplete Rules

⟨Event Description⟩
The March 28 Yichun Luming Mining tailings pond leakage accident was an environmental pollution incident...forcing 68,000 residents in Tieli City to temporarily switch to backup water sources. The incident resulted in direct economic losses of 44.2045 million yuan...

⟨Question⟩
According to the classification standards for sudden environmental incidents, events are generally categorized into 4 levels: extremely major, major, relatively major, and general. Given the wide impact range of this accident and the involvement of 68,000 affected people, it is determined to be an extraordinarily major environmental incident.

⟨Answer⟩
A
⟨Reasoning⟩
According to the provisions of the "National Emergency Response Plan for Sudden Environmental Incidents", sudden environmental incidents are classified into four levels based on their severity: extraordinarily major, major, relatively major, and general. One of the criteria for determining an extraordinarily major environmental incident (Level I) is any of the following: ...; requiring **the evacuation or relocation of more than 50,000 people** due to pollution; causing direct economic losses exceeding **100 million yuan**; resulting in loss of regional ecological functions or extinction of nationally protected species in the area; or causing interruption of water intake from centralized drinking water sources in prefecture-level or higher cities. Although this accident did not meet the casualty criteria, its wide impact involving 68,000 people temporarily switching to backup water sources, and the direct economic loss of 44.2045 million yuan, meet the criteria for classification as an extraordinarily major environmental incident.

(c) Fact Understanding Errors

Figure 12: Cases for error patterns.

lion RMB, this meets the criteria for an extra-large geological disaster risk and disaster level.

For extra-large geological disasters, the county-level people's government of the disaster area must, in accordance with the collective monitoring and prevention responsibility system, immediately notify the responsible parties for disaster prevention, monitoring personnel, and the local population within the geological hazard danger zone. It must make decisions on whether to evacuate the population and the emergency measures to be taken; promptly delineate the geological hazard danger zone; set up clear warning signs; determine alert signals and evacuation routes; organize population evacuation or implement risk mitigation and control measures; propose emergency responses based on the specifics of the risk and disaster situation; and, in critical situations, forcibly organize the evacuation of threatened residents.

Therefore, it is the county-level people's government that is responsible for delineating the hazard zone and organizing evacuations, that is Maoxian County People's Government.

⟨Event Description⟩

At 17:00 on March 15, 2024, a forest fire broke out in Baizi Village, Xiala Town, Yajiang County, Garze Prefecture, Sichuan Province. Influenced by extreme winds exceeding level 8, the fire reignited and rapidly spread, affecting 11 villages and 1 community, with a total of 837 households and 3,396 people impacted. The fire formed three separate fire zones at an average altitude of about 2,600 meters, with the fire lines spreading intermittently. Preliminary investigations indicate that the fire was caused by construction-related hot work encountering extreme winds, though the exact cause is still under investigation.

⟨Question⟩

The command should be led by the Forest and Grassland Fire Prevention and Control Headquarters of Garze Prefecture, Sichuan Province. Please evaluate whether this decision plan is appropriate.

⟨Answer⟩

A

⟨Reasoning⟩

According to the Forest Fire Prevention Regulations and the Working Rules of the Forest and Grassland Fire Prevention and Control Headquarters of Sichuan Province, emergency responses to forest fires should be commanded uniformly by the Forest and Grassland Fire Prevention and Control Headquarters. Since a forest fire occurred in Baizi Village, Xiala Town, Yajiang County, Garze Prefecture, Sichuan Province, the command should be the responsibility of the Forest and Grassland Fire Prevention and Control Headquarters of Garze Prefecture. Therefore, this decision plan is correct.

Figure 13: Case for reasoning and answers discrepancy.

⟨Event Description⟩

The full mountainside collapse on July 5, 2025, occurred at 13:25 in Longwei Village, Chengxiang Town, Tianquan County, Ya'an City, Sichuan Province, along National Highway 318. This natural disaster was triggered by continuous heavy rainfall, causing a landslide of approximately over 10,000 cubic meters of earth and rock. The landslide resulted in a 50-meter-long section of the highway being blocked, burying three trucks and one motorcycle. At the time of the accident, the area was under warnings for heavy rain, flash floods, and geological hazards. The incident caused five fatalities and two minor injuries....

⟨Question 1⟩

Geological disasters are classified into four levels based on their hazard severity and scale: extra-large, large, medium, and small geological disaster emergencies and disaster situations. According to relevant regulations, this disaster triggered a landslide of approximately over 10,000 cubic meters, indicating a severe situation. Therefore, it should be categorized a large-scale geological disaster.

⟨Answer 1⟩

A

⟨Reasoning 1⟩

According to the classification standards for geological disasters, the scale and severity of the hazard are key criteria for grading. In this incident, the landslide volume reached approximately over 10,000 cubic meters, causing the highway to be blocked, multiple vehicles to be buried, and casualties. This meets the definition of a large-scale geological disaster. Therefore, initiating a large-scale geological disaster is appropriate.

⟨Question 2⟩

Geological disasters are classified into four levels based on hazard severity and scale: extra-large, large, medium, and small geological disaster emergencies and disaster situations. According to relevant regulations, this disaster caused five fatalities and is categorized as a medium-scale geological disaster .

⟨Answer 2⟩

A

⟨Reasoning 2⟩

According to the classification standards for geological disasters, based on the severity and scale of the hazard, this disaster caused five fatalities and meets the definition of a medium-scale geological disaster emergency. Therefore, it is correct to classify this disaster as a medium-scale geological disaster.

Figure 14: Case for answer bias.

Category	Event Type	Definition
Natural Disasters	Forest Fire	An uncontrolled fire occurring in forested areas that causes damage to vegetation, wildlife, and potentially threatens human life and property.
	Geological Hazard	Geological processes that endanger human life and property, including mountain collapses, landslides, debris flows, ground subsidence, ground fissures, and land subsidence, which are triggered by natural factors or human activities.
	Flood and Drought	Sudden water and drought disasters include river floods and waterlogging, mountain floods (caused by rainfall-induced flash floods and debris flows), typhoon storm surges, droughts, water supply crises, and secondary disasters triggered by floods, storm surges, or earthquakes such as reservoir dam breaks, levee breaches, collapse of sluice gates, and formation of barrier lakes.
	Earthquake	Sudden shaking of the Earth's surface caused by the rapid release of energy in the Earth's crust, resulting in ground motion and potential damage.
Accident Incidents	Fire Incident	Uncontrolled combustion event that causes or may cause damage to life, property, or the environment.
	Environment Pollution	Events caused by pollutant discharge, natural disasters, or industrial accidents that result in the rapid release of toxic or harmful substances—such as pollutants or radioactive materials—into air, water, or soil, leading to environmental degradation, threats to public health and property, ecological damage, or significant social impact, requiring urgent response measures. These mainly include sudden air, water, soil pollution, and radiation pollution incidents.
	Coal Mine	Unexpected incidents occurring within coal mining operations that cause or may cause casualties, property damage, environmental harm, or disruption of mining activities, including explosions, collapses, gas leaks, and other hazardous events.
	Road Traffic	Urgent incidents caused by natural disasters, transport accidents, or socio-economic disruptions that result in or may result in major passenger hub interruptions, severe casualties, mass evacuations, environmental damage, serious social harm, or critical shortages in essential goods and passenger transport, requiring emergency response from transportation authorities.
Public Health	Infectious Diseases	Epidemics caused or potentially caused by legally designated infectious diseases, emerging infectious diseases, unexplained infectious diseases, and other major infectious diseases posing public health risks.
	Food Security	Incidents originating from food, such as foodborne diseases and food contamination, that cause or may cause harm to human health.
Society Security	Terrorist Attack	Acts of violence or intimidation intended to cause fear, harm public safety, or coerce authorities to achieve political or ideological goals.

Table 3: Predefined taxonomy.

Category	Event Type	EVENTS	TF	MCQ	MAQ	SAG	RG	Total
Natural Disasters	Forest Fire	20	52	30	10	22	5	119
	Geological Hazard	10	33	22	9	20	5	89
	Flood and Drought	10	40	26	11	19	5	101
	Earthquake	20	37	30	10	20	5	102
Accident Incidents	Fire Incident	11	40	33	11	16	5	105
	Environment Pollution	11	43	31	11	20	5	110
	Coal Mine	10	40	29	10	20	5	104
	Road Traffic	5	40	30	10	20	5	105
Public Health	Infectious Diseases	21	39	30	54	12	5	140
	Food Security	5	29	31	30	21	5	116
Society Security	Terrorist Attack	21	35	20	9	19	5	88
Total		144	428	312	175	209	55	1,179

Table 4: EDM-Bench Dataset Detailed Information.

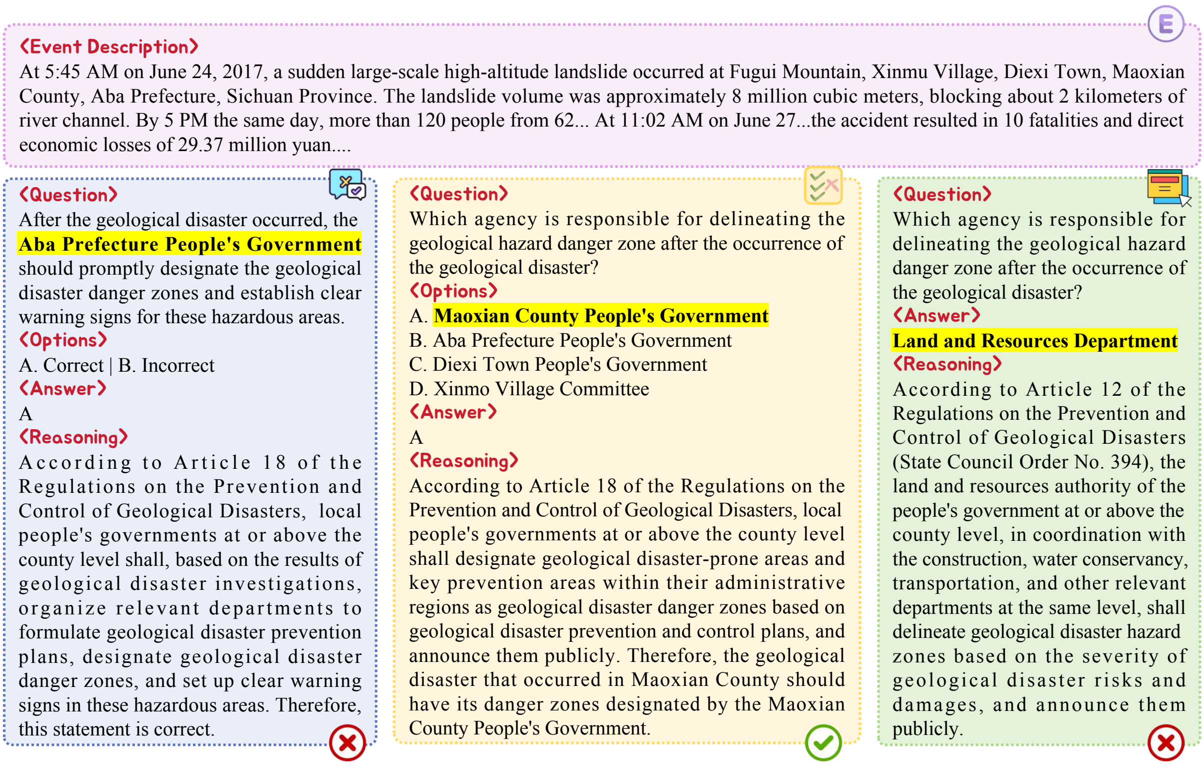

Figure 15: Case for sensitivity to task format.