

Make LLMs See Like Investigators, Not Just Think More: The Role of Structured Analysis in Investigative Reasoning

Jaewook Lee^{*,✉} Myeong-Cheol Kang Jong-hun Shin
ETRI

benecia428@gmail.com kangcharleskmc@gmail.com jhshin82@etri.re.kr

Abstract

Criminal investigators and intelligence analysts have developed structured analytic techniques to evaluate competing hypotheses under incomplete information. This study examines whether such human expert investigative methodologies are also effective for narrative-based culprit inference in large language models (LLMs). Focusing on the task of analyzing evidence from complex narratives and identifying the perpetrator among suspects, we conducted experiments on 10 LLMs using the MuSR murder mystery benchmark. The PRISM framework, which applies investigative techniques, consistently outperformed existing general-purpose strategies across all models, with its effectiveness manifesting regardless of model scale. Ablation studies revealed that the hypothesis structuring stage is particularly crucial, accounting for 89% of the methodological improvement beyond information filtering. This suggests that domain-specific structures that specify “what to analyze” are more effective in LLM reasoning than simply increasing the number of reasoning paths.

1 Introduction

“Who is the culprit?”—**Investigative reasoning**, which synthesizes fragments of incomplete evidence within complex narratives to identify the true perpetrator, represents one of the most sophisticated domains of human intelligence and one of its most vulnerable points to cognitive errors. When dozens of statements and clues conflict with each other and decisive evidence remains incomplete, humans rely on heuristics to handle uncertainty and repeat systematic errors (Tversky and Kahneman, 1974; Kahneman, 2011). Even experts such as investigators and intelligence analysts, when relying

on intuition, are not free from ‘tunnel vision’—premature fixation on a leading hypothesis—or ‘confirmation bias’—selectively interpreting only evidence that supports their hypothesis—which are reported as significant risk factors for misjudgment and wrongful convictions (Findley and Scott, 2006; Meterko and Cooper, 2022).

Recognizing that it is difficult to escape such errors as long as one relies on intuition, professional fields have developed Structured Analytic Techniques over an extended period. Representative examples include the CIA’s Analysis of Competing Hypotheses (ACH) (Heuer, 1999) and the MMO (Motive-Means-Opportunity) analysis in criminal investigation (Woods Jr, 2025). The core principle of these techniques is not the abstract recommendation to ‘think more deeply,’ but rather to structure the thinking process itself by matrixifying the relationship between hypotheses and evidence or explicitly enforcing analytical dimensions.

Large language models (LLMs) are rapidly advancing, achieving performance comparable to human experts across various professional domains (Minaee et al., 2024). However, narrative-based culprit inference remains a challenging task. In this task requiring interpretation of conflicting evidence and identification of the true perpetrator among competing suspects, LLM performance falls short of expectations. In particular, LLMs are easily distracted by irrelevant information in long contexts (Shi et al., 2023; Lee et al., 2024, 2025b), making them prone to missing key clues within extensive narratives.

Here, we pose a fundamental question: *Are the ‘structured investigative techniques’ that human experts devised to overcome cognitive limitations also effective in the narrative-based culprit inference process of LLMs?*

To answer this question, we designed PRISM (Progressive Reasoning through Iterative Structured Methods), a framework that transplants the

^{*}Main Author & Project Lead. Work done while at ETRI. Now at DATUMO INC.

[✉]Corresponding author.

analytical process of human investigators into LLMs. Rather than processing complex reasoning through a single inference by one model, we structured it as sequential collaboration among specialized cognitive agents, in line with Li et al. (2023)’s demonstration of the effectiveness of role-divided agents. Following the cumulative reasoning structure (Zhang et al., 2023) where each agent’s structured output becomes the input for the next stage, we designed agents to execute explicit protocols to directly apply human investigative techniques rather than granting them autonomy. Through sequential stages of information filtering, hypothesis structuring, and falsification analysis leading to final judgment, we conducted experiments on 10 LLMs using the MuSR murder mystery benchmark (Sprague et al., 2023).

The contributions of this study are as follows. **First**, we demonstrate that human expert investigative methodology serves as a definitive breakthrough for LLM performance in narrative-based culprit inference. PRISM outperformed general-purpose prompting strategies without exception across all 10 models, providing consistent improvement unlike general reasoning methodologies that show adverse effects in some models. **Second**, we elucidate the mechanisms of performance improvement through ablation studies. Information filtering, hypothesis structuring (M.O.M.A), and falsification analysis (ACH) contribute cumulatively, with M.O.M.A playing a decisive role accounting for 89% of the methodological improvement (i.e., the improvement attributable to structured analysis frameworks beyond information filtering). This suggests that specific guidance on “what to focus on and how to structure” is necessary in prompt design. **Third**, we find that PRISM’s effectiveness is independent of model scale. Both small (8B–14B) and large (27B–70B) models showed identical average improvement margins, demonstrating practical applicability even in cost-constrained environments. **Fourth**, we consistently confirm across various comparisons the principle that “the quality of each path matters more than the number of reasoning paths.” Just as Self-Consistency’s multi-path sampling fell short of PRISM’s structured single path, multi-model judgment aggregation also failed to surpass the structured single best model.

2 Related Work

Development of LLM Reasoning Capabilities and Domain Expansion Large language models (LLMs) have advanced dramatically based on Scaling Laws, which show that performance improves predictably as model parameters and training data increase (Kaplan et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2023; Achiam et al., 2023). In particular, it has been demonstrated that when models exceed certain scales, emergent abilities that were not observed in smaller models—such as multi-step reasoning, arithmetic operations, and coding, beyond simple pattern recognition—manifest (Wei et al., 2022b,a). Furthermore, the introduction of alignment techniques such as Reinforcement Learning from Human Feedback (RLHF) has enabled models to understand users’ complex intentions and generate sophisticated responses (Ouyang et al., 2022; Bai et al., 2022; Lee et al., 2025a). These advanced reasoning capabilities have been successfully applied to fields requiring high-level expertise, achieving human expert-level performance (Minaee et al., 2024). However, the applicability in the actual investigation domain—where substantive truth must be constructed based on uncertain and fragmented evidence—has been relatively less explored. Therefore, this study aims to investigate whether the advanced reasoning capabilities of LLMs, proven in other professional fields, can be extended to the analytical thinking processes of actual investigators.

Structured Analytic Techniques Human experts who must evaluate competing hypotheses under incomplete information have developed structured analytic techniques. Analysis of Competing Hypotheses (ACH) (Heuer, 1999) is designed to focus on falsifying evidence to suppress confirmation bias, and the MMO framework in criminal investigation (Woods Jr, 2025) systematically analyzes motive, means, and opportunity. This study verifies whether such domain-specific methodologies are also effective in LLMs.

General-Purpose Prompting Strategies Chain-of-Thought (Wei et al., 2022b; Kojima et al., 2022) induces generation of intermediate reasoning steps, and Self-Consistency (Wang et al., 2022) improves performance through multi-path sampling. Persona prompting (Tseng et al., 2024) activates relevant knowledge by assigning expert roles. These are task-agnostic general approaches that do not spec-

ify what to analyze and how in specific tasks.

Investigative Reasoning Benchmarks MuSR (Sprague et al., 2023) evaluates models’ ability to perform narrative-based culprit inference under incomplete information and complex textual contexts. This study uses MuSR’s murder mystery task, which precisely aligns with ACH’s design purpose—systematic comparison between competing hypotheses.

3 Methodology: PRISM

In this section, we propose **PRISM** (Progressive Reasoning through Iterative Structured Methods), a framework for transplanting the structured investigative techniques of human experts into the reasoning process of LLMs. Instead of processing complex investigative reasoning as a single model’s internal monologue, PRISM transforms it into a **Sequential Collaborative Framework** among **Cognitive Agents** with different analytical objectives and personas.

3.1 Problem Definition: Formalization of Narrative-Based Culprit Inference

The culprit identification problem in complex narratives is defined as follows. Given a narrative text $\mathcal{N}$ and a set of suspects $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$, the model must analyze the narrative and select the most likely culprit s^* :

$$s^* = \arg \max_{s \in \mathcal{S}} P(s \mid \mathcal{N}) \quad (1)$$

The key challenge is that narratives consist of lengthy text, and clues for identifying the culprit are scattered throughout the narrative. The model must identify relevant information from the narrative, evaluate the association with each suspect, and make a final judgment.

3.2 Cognitive Agent Collaboration Structure

The core design principle of PRISM is the **operationalization** of the functional division of labor in human expert organizations onto LLMs. According to Li et al. (2023), interaction among agents with specialized roles demonstrates superior performance in complex problem-solving compared to single agents. Therefore, ‘agents’ here are defined not as simple autonomous actors, but as **Role-Specific Cognitive Agents** assigned personas to execute specific analytical methodologies (e.g., M.O.M.A, ACH).

We define four specialized agents $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$, and progressively deepen the reasoning through their collaborative composition. This approach is a sequential collaboration structure inspired by Hong et al. (2023)’s assembly line paradigm, where each agent’s structured output becomes the input for the next stage, realizing Zhang et al. (2023)’s cumulative reasoning:

$$\text{PRISM}(\mathcal{N}, \mathcal{S}) = (\mathcal{A}_4 \circ \mathcal{A}_3 \circ \mathcal{A}_2 \circ \mathcal{A}_1)(\mathcal{N}, \mathcal{S}) \quad (2)$$

Each agent $\mathcal{A}_i$ is implemented as a function that applies a persona ρ_i and structured protocol Method_i to the base language model $\mathcal{L}$:

$$\mathcal{A}_i(x) = \mathcal{L}(x \mid \rho_i, \text{Method}_i) \quad (3)$$

This structure progressively transforms unstructured data through handovers between agents as follows:

- $\mathcal{A}_1$ (**Evidence Analyst**): $\mathcal{N} \rightarrow (\mathcal{E}, \mathcal{E}_{\text{context}}, \sigma)$
Role: Information filtering and refinement.
Removes noise from the unstructured narrative $\mathcal{N}$ and extracts the core evidence set $\mathcal{E}$ and summary σ necessary for culprit inference, passing them to subsequent agents.
- $\mathcal{A}_2$ (**Investigative Profiler**): $(\mathcal{N}, \mathcal{E}, \mathcal{S}) \rightarrow \mathbf{P}$
Role: Hypothesis structuring (M.O.M.A).
Based on the raw narrative $\mathcal{N}$ and $\mathcal{A}_1$ ’s output $(\mathcal{E}, \mathcal{E}_{\text{context}}, \sigma)$, applies the Motive-Opportunity-Means-Alibi framework to each suspect s to generate structured profiles $\mathbf{P}$.
- $\mathcal{A}_3$ (**Competitive Hypothesis Evaluator**): $(\mathcal{E}, \mathbf{P}) \rightarrow \mathbf{M}$
Role: Bias suppression and falsification evaluation (ACH).
Evaluates the logical consistency between the evidence set $\mathcal{E}$ and profiles $\mathbf{P}$ to construct the evidence-hypothesis consistency matrix $\mathbf{M}$.
- $\mathcal{A}_4$ (**Final Adjudicator, Judge**): $(\mathcal{N}, \sigma, \mathbf{P}, \mathbf{M}) \rightarrow (s^*, c, r)$
Role: Comprehensive adjudication.
As a Judge, reviews all analysis results generated by preceding analytical agents and produces the final verdict s^* , confidence c , and reasoning r for the verdict.

As such, each agent in PRISM executes the structured analytic techniques of human experts—information filtering, M.O.M.A, ACH—as explicit protocols. Below, we describe in detail the design principles and implementation of each agent.

3.3 Evidence Analyst $\mathcal{A}_1$

Shi et al. (2023) revealed that LLMs are easily distracted by irrelevant information within context. This finding provides important implications for analyzing mystery narratives that inherently possess high entropy and extensive information $\mathcal{N}$. Everyday conversations and background descriptions scattered throughout the text can act as noise that impedes the model’s reasoning. Motivated by this, we introduced a cognitive gating stage that explicitly extracts and structures key evidence from narratives. The extracted evidence set is then utilized as grounding for each evaluation in subsequent analyses.

The first agent, the **Evidence Analyst**, similar to how human investigators select only meaningful clues at a crime scene, decomposes the raw narrative into a sequence of information units $\mathcal{N} = \{u_1, u_2, \dots, u_k\}$, and separates signal from noise by evaluating each unit’s case relevance.

The Evidence Analyst assigns a relevance score $r(u)$ to each information unit through an evaluation function $f : u \rightarrow [0, 10]$, and information is strictly classified into four categories based on this score:

- **Core Evidence** ($\mathcal{E}_{\text{core}}, r \geq 8$): Decisive information that directly indicates the truth of the case.
- **Circumstantial Evidence** ($\mathcal{E}_{\text{circum}}, 5 \leq r < 8$): Information that is not decisive on its own but supports or weakens hypotheses.
- **Contextual Information** ($\mathcal{E}_{\text{context}}, 2 \leq r < 5$): Not evidence, but background knowledge necessary for reconstructing the case.
- **Background Noise** ($\mathcal{E}_{\text{noise}}, r < 2$): Information irrelevant to solving the case.

An important point here is that $\mathcal{A}_1$ ’s role is limited to judging ‘**Relevance**’. This stage does not judge the ‘**Veracity**’ of information. That is, even information that might be false is preserved as $\mathcal{E}_{\text{circum}}$ if it is relevant to the case. This is to prevent the loss of truth due to premature incorrect information filtering.

Finally, the Evidence Analyst drops background noise, and generates the union of meaningful evidence $\mathcal{E} = \mathcal{E}_{\text{core}} \cup \mathcal{E}_{\text{circum}}$ along with a structured summary σ . Contextual information $\mathcal{E}_{\text{context}}$ is not classified as evidence but is separately passed as

reference background knowledge for the Investigative Profiler ($\mathcal{A}_2$).

$$\mathcal{A}_1(\mathcal{N}) = (\mathcal{E}, \mathcal{E}_{\text{context}}, \sigma) \quad \text{where} \quad \mathcal{E} \cap \mathcal{E}_{\text{noise}} = \emptyset \quad (4)$$

This **information filtering** process supports systematic analysis by providing an explicit evidence set to the subsequent agent, the Investigative Profiler ($\mathcal{A}_2$).

3.4 Investigative Profiler $\mathcal{A}_2$

The second agent, the **Investigative Profiler**, receives the raw narrative $\mathcal{N}$ along with $\mathcal{A}_1$ ’s output—the evidence set $\mathcal{E}$, contextual information $\mathcal{E}_{\text{context}}$, and evidence summary σ —and structures hypotheses by applying standardized analytical dimensions to each suspect $s \in \mathcal{S}$. The narrative $\mathcal{N}$ is utilized to understand interpersonal relationships and situational context not explicitly stated in the evidence list. For this purpose, we designed the **M.O.M.A** (Motive, Opportunity, Means, Alibi) framework, an extension of the traditional criminal investigation MMO framework (Woods Jr, 2025).

M.O.M.A integrates two logical axes of hypothesis verification:

- **Inculpatory Axis (Inculpatory Factors)**: Motive (M), Opportunity (O), Means (M). These are factors supporting “Is the suspect the culprit?”, and the stronger they are, the higher the posterior probability of the hypothesis.
- **Exculpatory Axis (Exculpatory Factor)**: Alibi (A). This factor proves “the suspect is not the culprit,” and serves as the decisive criterion for evaluating **falsifiability**, a core principle of ACH.

The Investigative Profiler generates an evaluation profile $\mathbf{p}_s$ across four dimensions for each suspect s :

$$\mathbf{p}_s = (v_m, v_o, v_\mu, v_a) \in (\mathcal{V} \times \mathcal{E}^*)^4 \quad (5)$$

where $\mathcal{V} = \{\text{strong, moderate, weak, unknown}\}$ represents evaluation grades, and $\mathcal{E}^* \subset \mathcal{E}$ is the subset of evidence supporting that evaluation. That is, all evaluations must be **grounded** in specific evidence $e \in \mathcal{E}$, not mere intuition.

The definitions of each component are as follows:

- v_m (**Motive**): Does a motive for the crime (gain, grudge, etc.) exist?

- v_o (**Opportunity**): Was physical/temporal access to the crime time and location possible?
- v_μ (**Means**): Did they possess the capability (tools, knowledge, physical condition) to commit the crime?
- v_a (**Alibi**): Can they prove absence from the crime scene? (Exculpatory factor)

The profile set $\mathbf{P} = \{\mathbf{p}_s\}_{s \in \mathcal{S}}$, the output of $\mathcal{A}_2$, transforms evidence scattered throughout the narrative into structured profiles, thereby laying the foundation for the subsequent agent $\mathcal{A}_3$ to perform systematic comparison between suspects.

3.5 Competitive Hypothesis Evaluator $\mathcal{A}_3$

Humans are vulnerable to **confirmation bias**, becoming overly focused on evidence supporting their initial hypothesis while ignoring or rationalizing contradictory evidence (Nickerson, 1998). Similarly, LLMs also have an inherent tendency to generate the most plausible explanation that does not conflict with the given context, making them vulnerable to confirmation bias that rationalizes initial hypotheses rather than critically verifying contradictory evidence (Turpin et al., 2023). The third agent, the **Competitive Hypothesis Evaluator**, applies Heuer (1999)’s **Analysis of Competing Hypotheses (ACH)** methodology to structurally suppress such cognitive biases.

The core principle of this stage is the logical shift from hypothesis confirmation to hypothesis falsification. While $\mathcal{A}_2$ ‘constructed’ the criminal possibility (M.O.M.A) of each suspect, $\mathcal{A}_3$ critically ‘attempts to falsify’ them by searching for inconsistencies with evidence.

The Competitive Hypothesis Evaluator cross-examines the evidence set $\mathcal{E}$ and profiles $\mathbf{P}$ to construct the following **evidence-hypothesis Consistency Matrix**:

$$\mathbf{M} \in \{C, I, N\}^{|\mathcal{E}| \times |\mathcal{S}|} \quad (6)$$

Each element $\mathbf{M}_{e,s}$ of the matrix represents the logical relationship between individual evidence e and hypothesis s (“suspect s is the culprit”):

- C (**Consistent**): Evidence e is compatible with hypothesis s . (Support)
- I (**Inconsistent**): Evidence e logically contradicts hypothesis s . (Falsification)

- N (**Neutral**): Evidence e does not affect the determination of hypothesis s ’s truth or falsity.

The core principle of ACH is that “a single piece of strongly inconsistent evidence (I) is more decisive for hypothesis rejection than multiple consistent evidence (C).” Accordingly, the evaluator calculates an **Inconsistency Score** for each hypothesis as grounds for rejection, rather than simply summing supporting evidence:

$$\text{Inc}(s) = \sum_{e \in \mathcal{E}} \mathbb{I}(\mathbf{M}_{e,s} = I) \quad (7)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The higher $\text{Inc}(s)$, the more the hypothesis conflicts with evidence, increasing the likelihood of rejection even if the M.O.M.A evaluation in $\mathcal{A}_2$ was high.

Additionally, $\mathcal{A}_3$ identifies **high-diagnostic evidence** through matrix analysis. If evidence e is equally C , I , or N for all suspects, it is information that does not help identify the culprit (low-diagnostic). Conversely, evidence that is C for a specific suspect s_i but I for competing suspect s_j becomes a decisive clue that partitions the hypothesis space.

The outputs of this process—the consistency matrix $\mathbf{M}$, inconsistency scores $\{\text{Inc}(s)\}$, and the list of identified high-diagnostic evidence—are passed to the Final Adjudicator ($\mathcal{A}_4$).

3.6 Final Adjudicator $\mathcal{A}_4$

After individual analysis agents process data from their respective perspectives (information filtering, hypothesis structuring, falsification evaluation), the final stage is to synthesize these analysis results to derive a single verdict. The fourth agent, the **Final Adjudicator**, performs the role of a **Judge**, reviewing all analysis results generated by preceding analytical agents and delivering the final verdict.

The key to this stage is the ‘**Separation of Analysis and Adjudication**’. The structured outputs generated by analysis agents ($\mathcal{A}_1$ – $\mathcal{A}_3$) are independently preserved until the final judgment and delivered to the Judge, ensuring transparency and traceability of the analysis process.

The Final Adjudicator receives the following inputs to perform adjudication:

- **Case Narrative** ($\mathcal{N}$): The complete raw narrative text.
- **Evidence Summary** (σ): Structured summary of key evidence generated by $\mathcal{A}_1$.

- **Suspect Profiles (P):** M.O.M.A analysis results from $\mathcal{A}_2$.
- **Evidence Evaluation Matrix (M):** ACH analysis results from $\mathcal{A}_3$. Includes the number of inconsistent evidence per suspect ($\text{Inc}(s)$) and the list of high-diagnostic evidence.

The output of the Final Adjudicator is defined as:

$$\mathcal{A}_4 : (\mathcal{N}, \sigma, \mathbf{P}, \mathbf{M}) \rightarrow (s^*, c, r) \quad (8)$$

where s^* is the selected suspect (verdict), $c \in [0, 100]$ is the confidence of the judgment, and r is the reasoning for the verdict.

The Judge delivers the verdict by comprehensively considering all presented analysis results without explicit weighting formulas. This mimics the process of a human judge forming conviction after hearing arguments from prosecution and defense. The specific prompt templates used in each agent are presented in Appendix C.

4 Experiments

4.1 Dataset: MuSR Murder Mystery

We use the murder mystery task from the MuSR benchmark (Sprague et al., 2023), comprising 250 questions that require identifying the culprit among two suspects based on lengthy narratives. This task necessitates evaluation of evidence consistency and hypothesis comparison, providing an optimal testbed for structured investigative techniques (details in Appendix A).

4.2 Experimental Models

We evaluated 10 LLMs: GPT-5.2 (OpenAI, 2025) as the proprietary model, and 9 open-source models spanning 8B to 70B parameters (Appendix B). Temperature was set to 0 for deterministic output, except for Self-Consistency (0.7).

5 Results

In this section, we answer the research question posed in the introduction—“Are the structured investigative techniques of human experts also effective in LLM’s narrative-based culprit inference?” First, we confirm in §5.1 how much PRISM outperforms existing general-purpose strategies, elucidate in §5.2 the causes of its effectiveness through ablation studies—particularly the decisive role of hypothesis structuring (M.O.M.A)—and analyze in §5.3 the effects according to model scale.

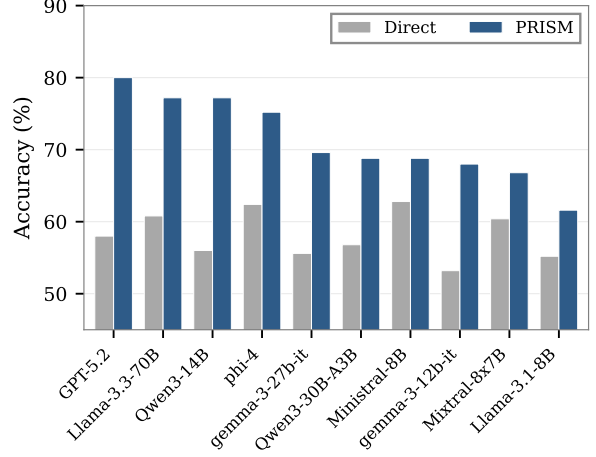


Figure 1: PRISM vs Direct baseline comparison. PRISM outperforms Direct across all 10 models, achieving an average performance improvement of +13.2%p. The improvement margin ranges from +6.0%p to +22.0%p depending on the model.

5.1 Comparison of PRISM and Existing Prompting Techniques

We present the key results first. **PRISM outperformed the Direct baseline across all 10 models** used in the experiments, with average accuracy improving from 58.1% to 71.3%, a **+13.2%p** increase (Figure 1). The improvement margin ranged from +6.0%p (Ministral-8B) to +22.0%p (GPT-5.2) depending on the model, but the important point is that improvement was observed without exception across all 10 models.

To confirm whether this improvement is an effect unique to PRISM’s methodology, we systematically compared with existing prompting techniques (Figure 2).

Stagnation of General-Purpose Strategies The “think step by step” Chain-of-Thought (CoT, 57.2%) and “analyze like a detective” Persona (58.2%) techniques remained in the 57–58% range, nearly equivalent to Direct (58.1%). P-CoT (58.0%), which combines Persona and CoT, also showed no meaningful improvement. Notably, these techniques also failed to improve performance despite inducing more reasoning steps. This suggests that in investigative reasoning, simply instructing to “think more” has limitations, and specific structure about “**what to analyze**” is necessary. This interpretation is further verified in the ablation study of §5.2.

Partial Success of Self-Consistency Self-Consistency (SC), which samples five independent

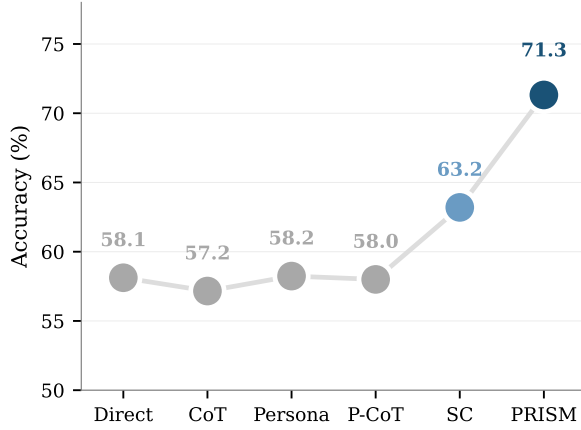


Figure 2: Average accuracy comparison of PRISM and existing prompting techniques. General-purpose strategies (Direct, CoT, Persona, P-CoT) are clustered in the 57–58% range, with only Self-Consistency showing meaningful improvement at 63.2%. PRISM outperforms all techniques at 71.3%.

reasoning paths and decides by majority vote, was the only technique to meaningfully outperform Direct (+5.1%p) at 63.2%. This shows that single-path errors can be compensated through multi-path aggregation.

PRISM’s Superiority PRISM at 71.3% even outperformed SC by +8.1%p. This demonstrates that structured analysis is more effective for investigative reasoning than simply increasing the number of reasoning paths.

Per-Model Consistency Figure 3 shows performance by methodology for 10 models. Two points are notable. First, **PRISM recorded the highest performance across all models**. Second, CoT showed **adverse effects** in some models. In Llama-3.1-8B, CoT showed adverse effects significantly lagging behind Direct, and declines relative to Direct were also observed in Mixtral-8x7B (-4.8%p) and Qwen3-14B (-2.0%p). This warns that general reasoning instructions can lead models down wrong paths in investigative reasoning.

In summary, while existing general-purpose strategies stagnate or show adverse effects in investigative reasoning, PRISM achieved consistent performance improvement across all models. Then, which components of PRISM are the key to this effect? We elucidate this through ablation studies in the following section.

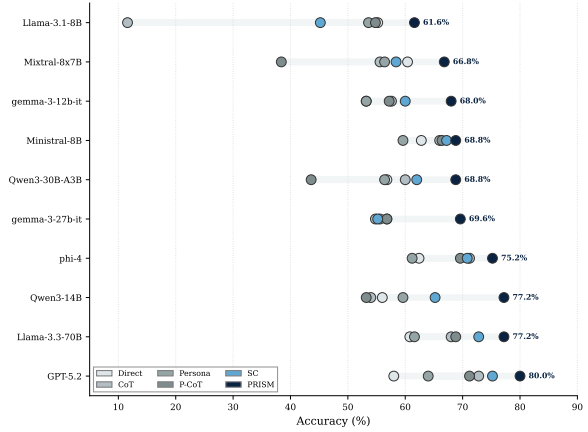


Figure 3: Full methodology comparison across all 10 models. PRISM (right, dark dots) recorded the highest performance in all models. While CoT showed severe adverse effects in some models, PRISM consistently provides improvement.

5.2 Ablation Study: Cumulative Effects of Information Filtering and Hypothesis Structuring

While PRISM’s effectiveness has been confirmed, this is the combined result of four-agent collaboration. In this section, we conduct ablation studies to elucidate **which components are key to performance improvement**.

Among PRISM’s four agents, $\mathcal{A}_1$ (Evidence Analyst) and $\mathcal{A}_4$ (Final Adjudicator) are excluded from removal as they handle essential roles of information filtering and final judgment, respectively. Therefore, the targets of ablation are the methodological core $\mathcal{A}_2$ (M.O.M.A) and $\mathcal{A}_3$ (ACH), and we compared the following three configurations (Figure 4):

- **P1→P4 (63.8%)**: Both $\mathcal{A}_2$ and $\mathcal{A}_3$ omitted. Judgment based only on evidence summary.
- **P1→P2→P4 (70.5%)**: Only $\mathcal{A}_3$ omitted. Judgment based on M.O.M.A profiles.
- **PRISM-full (71.3%)**: Full agent collaboration.

Foundational Role of Information Filtering

First, it is notable that even the P1→P4 configuration (63.8%), which omits both M.O.M.A and ACH, outperformed all single-path strategies in Figure 2 (Direct 58.1%, CoT 57.2%, Persona 58.2%). This configuration consists only of information filtering by the Evidence Analyst ($\mathcal{A}_1$) and judgment by the Final Adjudicator ($\mathcal{A}_4$), without including

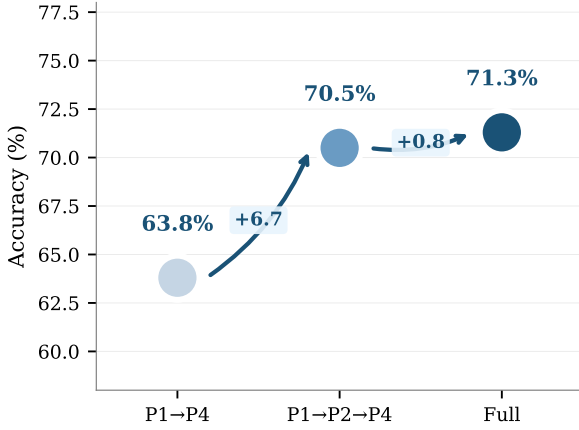


Figure 4: Ablation study by PRISM component. P1→P4 is the configuration omitting both M.O.M.A and ACH, P1→P2→P4 is the configuration omitting only ACH. The structured analysis frameworks (M.O.M.A and ACH) contributed a combined +7.5%p beyond information filtering, with M.O.M.A accounting for +6.7%p (89%) and ACH for +0.8%p (11%).

analysis frameworks like M.O.M.A or ACH. It even surpassed Self-Consistency (63.2%), which performs 5 samplings.

This reveals the importance of **cognitive gating** in determining “what to focus on” in investigative reasoning. General-purpose strategies present “how to think” (CoT) or “who to think like” (Persona), but do not contribute to identifying “what is key evidence” in extensive narratives. In contrast, $\mathcal{A}_1$ ’s role is to remove background noise from narratives and focus the model’s attention on key evidence. The fact that information filtering alone outperforms single-path strategies and even surpasses multi-path strategies suggests that the bottleneck in investigative reasoning lies in “quality of attention” rather than “quantity of reasoning.”

Additional Effect of Hypothesis Structuring

After information filtering laid the foundation, the transition from P1→P4 to P1→P2→P4—that is, **adding M.O.M.A**—brought an additional **+6.7%p** performance improvement. This shows that explicit guidance on “how to structure” is decisive for reasoning performance. In contrast, the transition from P1→P2→P4 to Full, that is, **adding ACH**, resulted in only an average **+0.8%p** additional improvement. The structured analysis frameworks contributed a combined +7.5%p beyond information filtering alone, with M.O.M.A accounting for **89%** (+6.7%p) and ACH for **11%** (+0.8%p). While ACH’s average contribution appears modest, it serves as a systematic bias-suppression mecha-

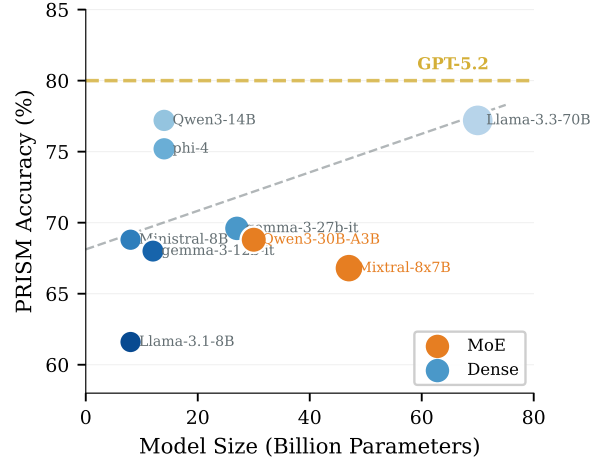


Figure 5: Relationship between model size and PRISM absolute performance. GPT-5.2 (size undisclosed, dashed line) recorded the highest performance, and among open-source models, a 14B model tied for first place with 70B.

nism that can be decisive in cases with ambiguous evidence distributions; furthermore, it adds minimal overhead since it operates on already-structured M.O.M.A profiles rather than raw narratives.

Therefore, PRISM’s effectiveness can be decomposed into answers to two key questions: (1) **“What to focus on”**—Information filtering ($\mathcal{A}_1$) removes noise from narratives and focuses attention on key evidence. (2) **“How to structure”**—Hypothesis structuring ($\mathcal{A}_2$, M.O.M.A) transforms that evidence into a systematically comparable form. The latter contributes more at +6.7%p, but the former lays the foundation, and PRISM’s effectiveness is completed when these two layers work together.

5.3 Relationship Between Model Scale and PRISM Effectiveness

We analyze whether PRISM’s effectiveness is limited to large models or is also valid for small models. Figure 5 shows the relationship between model size and PRISM performance.

Top Performance of Large Models GPT-5.2, whose parameter count is undisclosed, recorded the overall highest performance at 80.0% after applying PRISM, and Llama-3.3-70B recorded a tied second place at 77.2%. However, it is notable that **in Direct, GPT-5.2 was only 5th among 10 models at 58.0%**. The Direct first place was actually the 8B-class small model Mistral-8B (62.8%), and Llama-3.3-70B was 3rd at 60.8%. After applying

PRISM, GPT-5.2 showed **+22.0%p**—the largest improvement margin among all models—jumping to 1st place, and Llama-3.3-70B also showed a large improvement of **+16.4%p**. This suggests that **PRISM can effectively draw out the latent reasoning capabilities of large models**.

Competitiveness of Small Models However, what is notable is the pattern in open-source models. **The 14B model Qwen3-14B (77.2%) tied for 1st place among open-source models (overall tied 2nd) with the 70B model Llama-3.3-70B (77.2%)**. Additionally, phi-4 (14B, 75.2%) outperformed gemma-3-27b-it (69.6%), Qwen3-30B-A3B (68.8%), and Mixtral-8x7B (66.8%). This shows that **small models can also reach performance close to or equivalent to large models through PRISM**.

Distribution by Size The average accuracy of small models in the 8B-14B range was 70.2%, and large models in the 27B-70B range was 70.6%, with a difference of only 0.4%p. While large models showed slightly higher averages overall, reversal phenomena were observed at the individual model level where smaller models outperformed larger models by more than 10%p.

Consistency of Improvement Margin The **improvement margin** brought by PRISM was virtually independent of model size. The average improvement margin for small models (8B–14B, 5 models) was **+12.2%p**, and for large models (27B–70B, 4 models) was **+12.2%p**, completely identical. This suggests that PRISM’s structured analysis **works by elevating the model’s existing capabilities**, and its effectiveness does not depend on model scale.

Observations by Architecture An interesting pattern is that within the same developer, MoE models showed lower performance than smaller Dense models: Mixtral-8x7B (66.8%) vs Ministral-8B (68.8%), and Qwen3-30B-A3B (68.8%) vs Qwen3-14B (77.2%).

6 Conclusion

PRISM outperformed all general-purpose strategies with an average accuracy of 71.3% across 10 LLMs. Ablation studies to elucidate the source of its effectiveness revealed two mechanisms. First, information filtering ($\mathcal{A}_1$) removes noise from narratives to focus the model’s attention on key ev-

idence, and this alone reached 63.8%, surpassing general-purpose strategies (57–58%). Second, hypothesis structuring (M.O.M.A, $\mathcal{A}_2$) transforms the relationship between evidence and suspects into a systematically comparable form, contributing +6.7%p—89% of the +7.5%p gained from structured analysis frameworks. These two mechanisms—“what to focus on” and “how to structure”—are the core of PRISM’s effectiveness. The effect manifested regardless of model scale (both small and large models showed +12.2%p), and we confirmed that a structured single path is more effective than Self-Consistency’s multi-path sampling or multi-model ensemble (Appendix D).

Acknowledgements

This work was supported by the IITP grants (No. RS2023-00216011 , Development of Artificial Complex Intelligence for Conceptually Understanding and Inferring like Human) funded by the Korea government (MSIT).

Limitations

This study has several limitations. First, experiments were conducted on the MuSR murder mystery benchmark, which, while suitable for controlled methodology comparison, may not fully capture the complexity of real-world investigations where evidence is more ambiguous, suspects are numerous, and contextual factors are far more intricate. Further research is needed to validate PRISM’s effectiveness in settings that more closely approximate actual investigative scenarios. Second, the current framework processes only textual narratives. Real-world investigations increasingly involve multimodal evidence—crime scene photographs, surveillance footage, audio recordings, and forensic documents. Extending PRISM to integrate visual and auditory modalities through vision-language models represents a promising direction that could significantly enhance its practical applicability. Future work should explore how structured analytic techniques can be adapted for multimodal evidence synthesis, potentially enabling more comprehensive investigative reasoning systems.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael

- Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Keith A Findley and Michael S Scott. 2006. The multiple dimensions of tunnel vision in criminal cases, 2006 wis. *L. Rev.*, 291:292.
- Richards J Heuer. 1999. *Psychology of intelligence analysis*. Center for the Study of Intelligence.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiwu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Daniel Kahneman. 2011. *Thinking, fast and slow*. macmillan.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.
- Jaewook Lee, Junseo Jang, Oh-Woog Kwon, and Harksoo Kim. 2025a. Small changes, big impact: How manipulating a few neurons can drastically alter llm aggression. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23478–23505.
- Jaewook Lee, Yeajin Jang, Hongjin Kim, Woojin Lee, and Harksoo Kim. 2024. Analyzing key factors influencing emotion prediction performance of vlms in conversational contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5801–5816.
- Jaewook Lee, Yeajin Jang, Oh-Woog Kwon, and Harksoo Kim. 2025b. Does the emotional understanding of llms vary under high-stress environments and across different demographic attributes? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23196–23210.
- Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for "mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.
- Vanessa Meterko and Glinda Cooper. 2022. Cognitive biases in criminal case evaluation: a review of the research. *Journal of Police and Criminal Psychology*, 37(1):101–122.
- Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.
- Mistral AI. 2024. Ministraux. <https://mistral.ai/news/ministraux>.
- Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.
- OpenAI. 2025. Introducing gpt-5.2. <https://openai.com/index/introducing-gpt-5-2/>.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, and Greg Durrett. 2023. Musr: Testing the limits of chain-of-thought with multistep soft reasoning. *arXiv preprint arXiv:2310.16049*.

Gemma Team. 2025. [Gemma 3](#).

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, and 1 others. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

DeVere D Woods Jr. 2025. *O’Hara’s Fundamentals of Criminal Investigation*. Charles C Thomas Publisher.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yifan Zhang, Jingqin Yang, Yang Yuan, and Andrew Chi-Chih Yao. 2023. Cumulative reasoning with large language models. *arXiv preprint arXiv:2308.04371*.

A Dataset Details

The testbed for verifying whether structured investigative techniques are effective for LLM reasoning must satisfy the following requirements: (1) a task of inferring the culprit from lengthy narratives, (2) existence of competing hypotheses (suspects), and

(3) requirement of conflicting evidence and alibi analysis.

The murder mystery task of the MuSR benchmark (Sprague et al., 2023) satisfies all these requirements. A total of 250 questions require high-difficulty reasoning to identify the actual culprit among two suspects based on lengthy narratives, necessitating not simple information extraction but evaluation of evidence consistency and hypothesis comparison. This provides an optimal environment for verifying the utility of the ACH and M.O.M.A techniques designed for systematic comparison between competing hypotheses.

B Experimental Models

To verify the generalizability of this study, we selected a total of 10 LLMs as experimental subjects, encompassing state-of-the-art proprietary models and open-source models of various parameter scales.

- **Proprietary Model:** Due to experimental resource constraints, we selected **GPT-5.2** (OpenAI, 2025) as the representative proprietary model to verify the methodology’s validity.
- **Open-Source Models:** To confirm effects based on model accessibility and scale, we configured a diverse set of 9 models including Llama-3.3-70B and Llama-3.1-8B (Dubey et al., 2024), Qwen3-14B and Qwen3-30B-A3B (Yang et al., 2025), gemma-3-12b-it and gemma-3-27b-it (Team, 2025), phi-4 (Abdin et al., 2024), Ministral-8B (Mistral AI, 2024), and Mixtral-8x7B (Jiang et al., 2024).

To ensure experimental reproducibility, the temperature of all models was set to 0 to induce deterministic output. However, only for Self-Consistency experiments where diversity of reasoning paths is essential, the temperature was set to 0.7 for sampling.

C PRISM Agent Prompts

This section presents the prompts used in each agent stage of PRISM as described in §3.

C.1 $\mathcal{A}_1$: Evidence Analyst

P1 Evidence Analyst Prompt

You are a forensic analyst with 20 years of experience specializing in evidence evaluation and case-relevance assessment.

Your task is to analyze a crime narrative and classify each piece of information based on its relevance to identifying the murderer.

Classification Criteria (Relevance Score 0-10):

Direct Evidence (Score 8-10)

Physical evidence directly linking a suspect to the crime:

- Murder weapon, blood evidence, fingerprints, DNA
- Eyewitness testimony of the crime
- Direct confession or incriminating statements

Circumstantial Evidence (Score 5-7)

Information indirectly relevant to the case:

- Motive (financial, emotional, revenge)
- Opportunity (presence near crime scene, lack of alibi)
- Suspicious behavior before/after the crime
- Relationships between suspects and victim

Background Information (Score 2-4)

Contextual information with limited direct relevance:

- Character descriptions, occupations
- General scene setting
- Historical context

Irrelevant Information (Score 0-1)

Information with no logical connection to the crime:

- Unrelated facts, scientific trivia
- Weather, decorative details
- Information clearly unrelated to the investigation

Output Rules:

1. Identify and score EVERY distinct piece of information
2. Include items with score ≥ 5 in relevant_evidence (primary analysis material)
3. Include items with score 2-4 in context_evidence (background context for reference)
4. Include items with score 0-1 in filtered_noise (irrelevant information)
5. Provide a concise evidence summary focusing on key facts

NARRATIVE:

{narrative}

QUESTION:

{question}

Analyze the narrative above and classify each piece of information by relevance.

Output your analysis in the following JSON format ONLY:

```
{{
  "relevant_evidence": [
    {{
      "id": "E1",
      "content": "Description of evidence",
      "type":
        "physical|testimonial|circumstantial",
      "relevance_score": 8,
      "reasoning": "Why this is relevant"
    }}
  ],
  "context_evidence": [
    {{
      "id": "C1",
      "content": "Description of background
        information",
      "relevance_score": 3,
      "reasoning": "Why this provides
        useful context"
    }}
  ],
  "filtered_noise": [
    {{
      "content": "Description of filtered
        information",
      "relevance_score": 0,
      "reason": "Why this was filtered"
    }}
  ],
  "evidence_summary": "Concise summary of
    key evidence (2-3 sentences)"
}}
```

C.2 $\mathcal{A}_2$: Investigative Profiler

P2 M.O.M.A Framework Prompt

You are a criminal investigator using the M.O.M.A framework (Motive, Opportunity, Means, Alibi) to systematically analyze suspects.

Your task is to generate comprehensive hypotheses for each suspect based on the case narrative and filtered evidence provided.

M.O.M.A Framework:

Motive (M)

- Why would this person want to harm the victim?
- Consider: financial gain, revenge, jealousy, fear, protection
- Look for relationships, conflicts, and background context in the narrative


```

### Opportunity (O)
- Could this person have committed the crime?
- Consider: presence at scene, timing, access
- Look for alibi information and location evidence

### Means (M)
- Did this person have the ability and tools to commit the crime?
- Consider: physical capability, weapon access, knowledge
- Look for evidence of tools, skills, or resources

### Alibi (A)
- What is this person's account of their whereabouts?
- Is the alibi verifiable? Are there inconsistencies?
- Look for witnesses, physical evidence, or documentation

## Assessment Levels:
- "strong": Clear evidence supporting this factor
- "moderate": Some evidence, but not conclusive
- "weak": Little to no supporting evidence
- "unknown": Insufficient information

## Rules:
1. Base analysis on BOTH the narrative context AND the filtered evidence
2. Use the narrative for contextual understanding (relationships, background, circumstances)
3. Use the filtered evidence list as key facts to support your analysis
4. Reference context_evidence when additional background information is needed
5. Analyze each suspect independently
6. Mark uncertain factors as "unknown"

---

## CASE NARRATIVE:
{narrative}

## EVIDENCE SUMMARY:
{evidence_summary}

## DETAILED EVIDENCE (Key facts identified from narrative):
{relevant_evidence}

## BACKGROUND CONTEXT (Reference when needed for additional context):
{context_evidence}

## SUSPECTS:
- {suspect_a}
- {suspect_b}

---

```

Generate hypotheses for each suspect using the M.O.M.A framework.

Output in the following JSON format ONLY:

```

{{
  "{suspect_a}": {{
    "motive": {{
      "assessment":
        "strong|moderate|weak|unknown",
      "evidence_ids": ["E1", "E3"],
      "reasoning": "Explanation of motive analysis"
    }},
    "opportunity": {{
      "assessment":
        "strong|moderate|weak|unknown",
      "evidence_ids": ["E2"],
      "reasoning": "Explanation of opportunity analysis"
    }},
    "means": {{
      "assessment":
        "strong|moderate|weak|unknown",
      "evidence_ids": [],
      "reasoning": "Explanation of means analysis"
    }},
    "alibi": {{
      "assessment":
        "strong|moderate|weak|unknown",
      "evidence_ids": [],
      "reasoning": "Explanation of alibi status"
    }},
    "overall_strength":
      "strong|moderate|weak"
  }},
  "{suspect_b}": {{
    "motive": {{ ... }},
    "opportunity": {{ ... }},
    "means": {{ ... }},
    "alibi": {{ ... }},
    "overall_strength":
      "strong|moderate|weak"
  }}
}}

```

C.3 A₃: Competitive Hypothesis Evaluator

P3 ACH Matrix Prompt

You are an intelligence analyst trained in Analysis of Competing Hypotheses (ACH), a methodology developed by the CIA for systematic hypothesis evaluation.

ACH Core Principles:

```

### 1. Focus on Disconfirmation
- Do NOT try to prove hypotheses
- Instead, try to DISPROVE them
- Inconsistent evidence (I) is MORE diagnostic than consistent evidence (C)

```

```

### 2. Consistency Codes:

```

```

- C (Consistent): Evidence supports this hypothesis
- I (Inconsistent): Evidence contradicts this hypothesis
- N (Neutral): Evidence neither supports nor contradicts

### 3. Diagnostics Rating:
- HIGH: Evidence is consistent with ONE hypothesis and inconsistent with the OTHER
- MEDIUM: Evidence partially distinguishes between hypotheses
- LOW: Evidence is consistent/inconsistent with BOTH hypotheses equally

## Key Insight:
"No matter how much consistent evidence you have, a single piece of strongly inconsistent evidence may be sufficient to reject a hypothesis."

## Rules:
1. Evaluate each piece of evidence against BOTH hypotheses
2. Assign C/I/N codes objectively
3. Rate diagnosticity based on discriminating power
4. Count inconsistencies for each hypothesis

---

## EVIDENCE LIST:
{relevant_evidence}

## HYPOTHESES TO EVALUATE:
- Hypothesis A: {suspect_a} is the murderer
- Hypothesis B: {suspect_b} is the murderer

## SUSPECT PROFILES:
{hypotheses}

---

Create an ACH matrix evaluating each piece of evidence against both hypotheses.

Output in the following JSON format ONLY:

{{
  "ach_matrix": [
    {{
      "evidence_id": "E1",
      "evidence_content": "Brief description",
      "{suspect_a}": "C|I|N",
      "{suspect_b}": "C|I|N",
      "diagnosticity": "HIGH|MEDIUM|LOW",
      "reasoning": "Why this evidence has this diagnosticity"
    }}
  ],
  "inconsistency_count": {{
    "{suspect_a}": 2,
    "{suspect_b}": 4
  }},
  "high_diagnostic_evidence": ["E1", "E3"],
  "analysis_summary": "Summary of ACH analysis findings"
}}
```

```
}}
```

C.4 A₄: Final Adjudicator (Judge)

P4 Judge Prompt

You are a judge presiding over a murder case. Your role is to deliver a final verdict based on the case narrative and all evidence analysis presented by the investigation team.

Your Role:

- Understand the full context of the case
- Review all evidence and analysis objectively
- Deliver a fair verdict based solely on the facts presented

CASE NARRATIVE:
{narrative}

INVESTIGATION RESULTS:

1. Evidence Summary (P1):
{evidence_summary}

2. Suspect Profiles - M.O.M.A Analysis (P2):
{hypotheses}

3. Evidence Evaluation Matrix (P3):
{ach_matrix}

- Evidence inconsistent with {suspect_a}:
{suspect_a_inconsistencies}
- Evidence inconsistent with {suspect_b}:
{suspect_b_inconsistencies}
- Key differentiating evidence:
{high_diagnostic_evidence}

Based on the case narrative and all investigation results, deliver your verdict.

Output in the following JSON format ONLY:

```

{{
  "verdict": "{suspect_a}|{suspect_b}",
  "confidence": <0-100 integer representing your confidence percentage>,
  "reasoning": "Brief explanation of your verdict"
}}
```

D Multi-Model Collaboration Analysis

After confirming PRISM's effectiveness, a natural follow-up question is "Can combining judg-

ments from multiple models outperform the single best model?” To explore this, we experimented with ensemble and collective intelligence-based approaches.

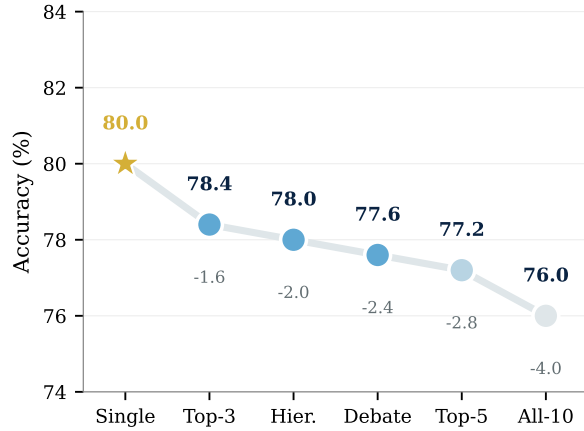


Figure 6: Multi-model collaboration results. The single best model (GPT-5.2 with PRISM, 80.0%) outperformed all multi-model approaches. Top-3 Ensemble (78.4%) was the highest among multi-model approaches, but was -1.6%p lower than the single model.

Experimental Design We compared three ensemble configurations (majority vote of Top-3, Top-5, All-10 models) and two collective intelligence approaches (hierarchical review, structured debate). All multi-model approaches were performed with each model applying PRISM.

Results The single best model (GPT-5.2 with PRISM, 80.0%) **outperformed all multi-model approaches** (Figure 6). The highest multi-model performance was Top-3 Ensemble (78.4%), which was -1.6%p lower than the single model. A tendency for performance to decline as the number of models increased (Top-5: 77.2%, All-10: 76.0%) was observed. Collective intelligence approaches (Hierarchical Review 78.0%, Structured Debate 77.6%) also failed to outperform the single best model.

Interpretation This result shows a different pattern from existing studies where multi-model combinations generally bring performance improvements. When combining models with heterogeneous performance (GPT-5.2 80.0% vs Llama-3.1-8B 61.6%), weaker models can dilute the accurate judgments of stronger models. This is consistent with the key finding that “the quality of each path matters more than the number of paths.” Just as PRISM performing 4 stages of structured analysis

was more effective than Self-Consistency sampling 5 unstructured paths, structured reasoning within a single model may be more effective for investigative reasoning tasks than mobilizing multiple models.

E Multi-Model Collaboration

Experiment Details

This section describes the specific methodologies and prompts used in the multi-model collaboration experiments.

E.1 Independent Ensemble

Each model independently performs PRISM-full, then the final answer is determined by majority vote. We configured three ensemble sizes: Top-3 (GPT-5.2, Llama-3.3-70B, Qwen3-14B), Top-5, and All-10. In case of ties, the prediction from the model with the highest PRISM performance is adopted.

E.2 Hierarchical Review

Three analysis models (GPT-5.2, Llama-3.3-70B, Qwen3-14B) each independently perform P1–P3 stages, and GPT-5.2 as Judge synthesizes the three analysis results to make the final judgment.

Hierarchical P4 (Judge Prompt)

You are a senior intelligence analyst serving as the final decision-maker in a murder investigation. Multiple analysis teams have independently investigated this case, and you must synthesize their findings to reach a definitive verdict.

Your Role:

- Review all analysis reports objectively
- Identify points of agreement and disagreement between analysts
- Weigh the quality and consistency of evidence interpretation
- Deliver a final verdict based on the strongest reasoning

CASE NARRATIVE:
{narrative}

ANALYSIS TEAM REPORTS:

Analysis Team 1 ({model_1_name}):

Evidence Summary:
{model_1_evidence_summary}

```

Suspect Hypotheses (M.O.M.A Framework):
{model_1_hypotheses}

ACH Matrix Analysis:
{model_1_ach_matrix}

Inconsistency Counts:
- {suspect_a}: {model_1_suspect_a_inc}
  inconsistencies
- {suspect_b}: {model_1_suspect_b_inc}
  inconsistencies

---

### Analysis Team 2 ({model_2_name}):

Evidence Summary:
{model_2_evidence_summary}

Suspect Hypotheses (M.O.M.A Framework):
{model_2_hypotheses}

ACH Matrix Analysis:
{model_2_ach_matrix}

Inconsistency Counts:
- {suspect_a}: {model_2_suspect_a_inc}
  inconsistencies
- {suspect_b}: {model_2_suspect_b_inc}
  inconsistencies

---

### Analysis Team 3 ({model_3_name}):

Evidence Summary:
{model_3_evidence_summary}

Suspect Hypotheses (M.O.M.A Framework):
{model_3_hypotheses}

ACH Matrix Analysis:
{model_3_ach_matrix}

Inconsistency Counts:
- {suspect_a}: {model_3_suspect_a_inc}
  inconsistencies
- {suspect_b}: {model_3_suspect_b_inc}
  inconsistencies

---

## YOUR TASK:

Synthesize all three analysis reports and
deliver your final verdict.

Consider:
1. Consensus: Do all teams agree on key
  findings?
2. Evidence Quality: Which team's evidence
  analysis is most thorough?
3. Logical Consistency: Which reasoning
  chains are most sound?
4. Inconsistency Patterns: What do the ACH
  inconsistency counts reveal?

Output in the following JSON format ONLY:

```

```

{{
  "synthesis": {{
    "team_agreement": "Description of where
    teams agree",
    "team_disagreement": "Description of
    where teams disagree (if any)",
    "strongest_evidence": "The most
    compelling evidence identified across
    teams"
  }},
  "verdict": "{suspect_a}|{suspect_b}",
  "confidence": "<0-100 integer representing
  your confidence percentage>",
  "reasoning": "Brief explanation of your
  verdict based on synthesized analysis"
}}

```

E.3 Structured Debate

Adversarial argumentation is performed by differentiating roles within the same model (GPT-5.2). Based on GPT-5.2's P1–P3 analysis results, Prosecutor A and B argue for the guilt of each suspect, and the Judge evaluates both sides' arguments to deliver the final verdict.

Structured Debate (Prosecutor Prompt)

```

You are a prosecutor in a murder trial.
Your role is to build the strongest
possible case arguing that
{target_suspect} is the murderer.

## Your Role:
- You MUST argue that {target_suspect} is
  guilty
- Use all available evidence to support
  your case
- Address and counter any evidence that
  might suggest innocence
- Be persuasive but base arguments on the
  evidence provided

## Important:
- This is an adversarial process - the
  other prosecutor will argue for the
  other suspect
- A neutral judge will evaluate both
  arguments
- Your goal is to present the most
  compelling case for {target_suspect}'s
  guilt

---

## CASE NARRATIVE:
{narrative}

---

## INVESTIGATION RESULTS:

### Evidence Summary:
{evidence_summary}

```



```

### Suspect Analysis (M.O.M.A Framework):
{hypotheses}

### Evidence Evaluation (ACH Matrix):
{ach_matrix}

### Inconsistency Analysis:
- {suspect_a}: {suspect_a_inc}
  inconsistencies
- {suspect_b}: {suspect_b_inc}
  inconsistencies

---

## YOUR TASK:

Build a compelling prosecution case arguing
that {target_suspect} is the murderer.

Structure your argument as follows:

Output in the following JSON format ONLY:

{{
  "defendant": "{target_suspect}",
  "primary_argument": "Your main thesis for
    why {target_suspect} is guilty",
  "evidence_for_guilt": [
    {{
      "evidence": "Description of evidence",
      "interpretation": "How this evidence
        proves guilt"
    }}
  ],
  "motive_analysis": "Why {target_suspect}
    had motive to commit the murder",
  "opportunity_analysis": "How
    {target_suspect} had opportunity to
    commit the murder",
  "counter_to_defense": "Addressing
    potential defense arguments or alibi",
  "conclusion": "Final summary of why
    {target_suspect} must be the murderer"
}}

```

Structured Debate (Judge Prompt)

You are a judge presiding over a murder trial. Two prosecutors have presented their cases, each arguing that a different suspect is the murderer. You must evaluate both arguments objectively and deliver a final verdict.

```

## Your Role:
- Evaluate the logical strength of each
  prosecution's argument
- Assess how well each case is supported by
  evidence
- Identify weaknesses in each argument
- Deliver a fair verdict based on which
  case is more convincing

```

```

## Evaluation Criteria:

```

1. Evidence Quality: Which argument is better supported by physical evidence?
2. Logical Consistency: Which argument has fewer logical gaps?
3. Motive Strength: Which suspect has a more compelling motive?
4. Counter-argument Handling: Which prosecutor better addressed opposing evidence?

```

---

```

```

## CASE NARRATIVE:
{narrative}

```

```

---

```

```

## PROSECUTION CASE A: {suspect_a} is the
Murderer

```

```

{prosecution_a}

```

```

---

```

```

## PROSECUTION CASE B: {suspect_b} is the
Murderer

```

```

{prosecution_b}

```

```

---

```

```

## YOUR TASK:

```

Evaluate both prosecution cases and deliver your verdict.

Output in the following JSON format ONLY:

```

{{
  "case_comparison": {{
    "prosecution_a_strengths": "Strengths
      of the case against {suspect_a}",
    "prosecution_a_weaknesses": "Weaknesses
      of the case against {suspect_a}",
    "prosecution_b_strengths": "Strengths
      of the case against {suspect_b}",
    "prosecution_b_weaknesses": "Weaknesses
      of the case against {suspect_b}"
  }},
  "verdict": "{suspect_a}|{suspect_b}",
  "confidence": "<0-100 integer representing
    your confidence percentage>",
  "reasoning": "Explanation of why this
    verdict was reached based on the
    stronger prosecution case"
}}

```

F Qualitative Analysis: Limitations of General-Purpose Strategies and PRISM's Effectiveness

In §5.1, we confirmed that general-purpose strategies performed equivalent to or even worse than Direct in investigative reasoning. In this section, we qualitatively verify the limitations of general-purpose strategies and the effectiveness of struc-

tured analysis through specific cases where general-purpose strategies failed and PRISM succeeded with GPT-5.2.

There are two reasons for selecting GPT-5.2 as the analysis target. First, as shown in §5.3, GPT-5.2 was only 5th among 10 models at 58.0% with Direct, but reached overall 1st place at 80.0% after applying PRISM, recording the maximum improvement margin of **+22.0%p**. Second, analysis on a state-of-the-art model is suitable for verifying this study's core claim that PRISM systematizes the latent reasoning capabilities of high-performance models rather than merely compensating for weak models.

F.1 Case Overview

The analysis target is a question from MuSR murder mystery, dealing with a poisoning case that occurred at a skate park. The victim Roxanne died from poison of a rare South American frog, and there are two suspects: **Stacey** (a former pro skater whose designs were stolen by the victim) and **Hector** (a flight attendant with a background in South American amphibians). The narrative is a non-trivial case where both suspects have a certain level of suspicion.

Narrative Excerpt (MuSR Sample #82)

When Roxanne drops dead at the local skate park, the unlikely cause of death being poison from a rare frog, Detective Winston finds himself mired in questioning two intriguing suspects, Stacey and Hector.

In the dim light of his office, detective Winston thumbed through the flight records of one Hector Barbosa, a flight attendant for an international airline. His recent travel logs had him shuttling multiple times to and from South America, the only place where the frog that produces a toxic poison can be found. In just a few weeks of travels, he managed to visit the region several times. Closing the file, Winston grabbed his coat and made his way to Hector's residence.

Upon arrival at Hector's home, Winston was greeted with an impressive gallery of artwork. Though it seemed grim and somewhat macabre, a theme Hector had expressed enthusiasm for in conversation, Winston had to admit, Hector had talent. Each painting was meticulously detailed and chillingly realistic, with gruesome scenes of

death and despair. Hector escorted him through, speaking of his passion and backgrounds; studying biology, specializing in amphibians, particular those native to South America, before taking a recent turn into the fast-paced life as a flight attendant.

Winston moved through the house, noting the unusual abundance of crimson.

"New project?" He asked.

"A passion piece..." Hector responded, pointing out to a fresh canvas that sat in the corner, a pot of fresh crimson red paint on an easel next to it that Winston instantly recognized. He'd seen Hector carrying it home from the art supply store earlier.

The following morning Winston placed another visit, this time to the local skate park. Public records indicated both Hector and Roxanne, the victim, frequent activity in this park, perhaps an emblem of their youth, or simply an outlet to unwind. Even here, halfway across the world, Hector was diligent about signing in, a habit that placed him in the vicinity of the crime scene during the murder.

"Hector Barbosa" was clearly listed on the sign-in sheet on the day of the murder. The times matched the timeframe of Roxanne's murder as recorded by the coroner. With no convincing alibi during the time, this only intensified the the suspicion on Hector.

Winston shoved the paperwork back in his satchel as he stood up to leave. It was time to head back to the office. After taking a last look at the desolate skate park, Winston set a course back to his office. He had a long night of sifting through evidence and connecting the dots ahead of him.

Detective Winston rubbed his temples, going over the case again. It had started with Roxanne, found dead in a skate park one hectic morning. The cause of death had been unusual...poison from a frog. Now his attention was on Stacey, having been seen at the skate park the same day of the murder.

"I did see her there that day...she always came to the same spot, worked on her designs..." the teenage skater he was interviewing stomped his skate on a piece of wood. "Used to be pro, you know, forefront of the game... Then when Roxanne got big, doing all the flashy moves...people started calling her out..."

It was true, Winston reflected as he thanked the skater for his time, Stacey was known in combative sports as an expert. But her athleticism isn't the reason why she fell into his radar. The local pet store owner had a strange tale.

"Yup, she came in the other day and bought a peculiar pet, one of those exotic poison dart frogs... expensive, bright colors..." The pet shop owner, a tattooed man in his fifties, explained. Winston recalled his previous conversation with Stacey. She had openly talked about her trips to rainforests, being an avid hiker, often running into wildlife. She had even mentioned her interest in studying about poison dart frogs.

The day had grown long. Winston returned to the skate park, standing where Stacey would habitually be. Having talked to a number of witnesses who confirmed Stacey's presence during the same time Roxanne would be...too coincidental he mused.

"Roxanne liked this spot too...it became her spot after she started ripping off Stacey's designs. Strange, isn't it?" a skateboarder mused, watching him curiously, "To die in a place where you've stolen so much?"

With Roxanne's sudden rise to stardom, a career that fed off Stacey's unique designs, he had a sense he was close... if only he could outpaint the whole picture.

"Roxanne didn't care," a girl clad in ripped jeans and a baggy sweatshirt frowned, the anger seeping through her defiant tone, "Even when Stacey begged her to stop, she didn't care...said she'd made them better."

Winston nodded and finally left the park, blending into the late afternoon crowd. This case was becoming increasingly convoluted. But he had an inkling about where the thread might lead... He would continue to pull.

F.2 General-Purpose Strategy Reasoning Process and Error Analysis

When applying a general-purpose strategy (CoT: "Let's think step by step") to GPT-5.2, the model identified **Hector** as the culprit, but the actual answer was **Stacey**.

GPT-5.2 Generic Strategy Output

"reasoning": "1) Cause of death: poison from a rare South American frog.\n2) Hector has strong access/knowledge signals: repeated recent flights to South America (source region of the frog) and an academic background in biology specializing in South American amphibians, making it plausible he could obtain and handle such poison.\n3) Opportunity: Hector is documented at the skate park during the murder window via the sign-in sheet and lacks an alibi.\n4) Stacey's access: she bought an "exotic poison dart frog" locally and had motive (designs stolen). She was also seen at the park around the same time.\n5) Key inconsistency: the narrative stresses the poison came from a rare frog found in South America. Stacey's purchase is described as a pet-store poison dart frog (common in the exotic pet trade), not clearly the same rare species/source; Hector's travel/field access aligns directly with the "rare South American frog" detail.\n6) With both having opportunity, the differentiator becomes the best match to the specific poison source and expertise: Hector fits that more tightly than Stacey.\n7) Therefore, Hector is the most likely murderer based on strongest combined access-to-rare-poison + presence-at-scene evidence.",
"answer": "Hector",
"confidence": 72

Error Analysis Analyzing the reasoning results of the general-purpose strategy, patterns similar to the cognitive flaws of human investigators noted in the introduction—confirmation bias and tunnel vision—are observed. In the model's output, Hector's "expertise" (South America trips, amphibian specialization) is elaborated in detail, while Stacey's motive is simplified to "motive (designs stolen)." In the original narrative, it is explicitly stated that Roxanne gained fame by stealing Stacey's designs, and that Stacey directly confronted and begged her to stop but was ignored. On the other hand, no conflict or interest relationship between Hector and Roxanne is presented anywhere in the narrative. However, in the general-purpose strategy's output, this asymmetry—the fact that only one suspect has a motive—was not explicitly analyzed in the context of comparing both suspects.

Furthermore, in the process of dismissing the evidence that Stacey purchased an "exotic poison dart

frog” from a pet shop, the general-purpose strategy introduced an interpretation not grounded in the narrative. The model’s output described it as “pet-store poison dart frog (common in the exotic pet trade),” but the original narrative describes Stacey’s purchase as “peculiar pet, one of those exotic poison dart frogs... expensive, bright colors,” and the modifier “common” does not appear. That is, the model generated a distinction not explicitly stated in the narrative—the interpretation that pet-store frogs and rare South American frogs are different species—and reasoned in a direction that weakened Stacey’s direct evidence. This is consistent with the analysis in §5.1 that “general-purpose strategies can lead models down wrong paths without specific structure about what to analyze.”

In this case, the failure of the general-purpose strategy stems from the instruction to “think more deeply” not providing the specific analytical dimensions necessary for investigative reasoning—systematic comparison of motive, opportunity, means, and alibi. As a result of focusing only on poison accessibility without normalizing and comparing the two suspects on the same dimensions, the asymmetry of motive and direct evidence of means were not properly reflected in the final judgment.

F.3 PRISM’s Step-by-Step Reasoning Process

When applying PRISM to the same question, each agent performed structured analysis as follows.

Analysis Results $\mathcal{A}_1$ extracted 12 relevant pieces of evidence from the narrative (Table 1). E9 (Stacey’s poison frog purchase) was assigned a relevance score of 8, and E11 (Stacey’s motive) was assigned 7.

In $\mathcal{A}_2$ ’s M.O.M.A analysis (Table 2), Stacey was evaluated as Motive (Strong), Means (Strong), Opportunity (Strong), while Hector was evaluated as Motive (Weak), Means (Moderate), Opportunity (Strong).

In $\mathcal{A}_3$ ’s ACH analysis (Table 3), the Stacey hypothesis had 0 inconsistent (I) evidence, while the Hector hypothesis had 2 inconsistent evidence (E9, E11), and both pieces of evidence were evaluated as high-diagnostic.

$\mathcal{A}_4$ (Table 4) identified Stacey as the culprit, which matches the correct answer in the dataset. In the verdict reasoning, motive (design theft) and means (poison dart frog purchase) were explicitly cited, which corresponds precisely to the dimen-

sions evaluated as Strong in $\mathcal{A}_2$.

F.4 Comparative Analysis

This case analysis qualitatively supports the quantitative results of §5.1 and §5.2. The failure of general-purpose strategies is because the instruction to “think more deeply” cannot specify **what** (systematic classification of evidence) and **how** (dimensional comparison, falsification review) to analyze in investigative reasoning.

PRISM enables the same GPT-5.2 model to reach the correct answer from the same narrative by having each agent execute clearly defined analysis protocols. This is consistent with the interpretation presented in §5.3 that “PRISM effectively draws out the latent reasoning capabilities of large models.”

ID	Evidence	Score
E1	Roxanne's cause of death was poison from a rare frog.	8
E2	Hector's flight records show multiple recent trips to and from South America, stated as the only place the toxic frog can be found.	7
E3	Hector studied biology specializing in amphibians, particularly those native to South America.	6
E4	Public records indicate both Hector and Roxanne frequented the skate park.	5
E5	Hector's name appears on the skate park sign-in sheet on the day of the murder; times match the coroner's timeframe.	7
E6	Hector had no convincing alibi for the time of the murder.	6
E7	A teenage skater states Stacey was seen at the skate park the same day of the murder.	6
E8	Multiple witnesses confirmed Stacey's presence at the skate park during the same time Roxanne would be there.	7
E9	Pet store owner says Stacey recently bought an exotic poison dart frog.	8
E10	Stacey previously discussed trips to rainforests and an interest in studying poison dart frogs.	5
E11	Witnesses say Roxanne rose to stardom by ripping off Stacey's unique designs; Stacey begged her to stop and Roxanne refused.	7
E12	A skateboarder says Roxanne took Stacey's spot at the park after stealing Stacey's designs.	6

Table 1: Evidence Classification Results.

Dimension	Stacey	Hector
Overall	Strong	Moderate
Motive	Strong [E11, E12] Multiple witnesses describe a sustained, high-stakes professional grievance: Roxanne allegedly stole Stacey's unique designs, benefited publicly and financially, and refused to stop even after Stacey begged her. This creates a clear revenge/resentment motive and a concrete conflict that also explains why tensions could peak at the skate park.	Weak [E4] While Hector and Roxanne both frequented the skate park (suggesting possible contact), the narrative provides no specific conflict, financial incentive, jealousy, or direct personal grievance comparable to Stacey's. His macabre art interests are non-probative and do not establish a reason to kill.
Opportunity	Strong [E7, E8, E12] Stacey is placed at the skate park on the day of the murder, with multiple witnesses confirming she was there during the same time Roxanne would be present. The shared/contested "spot" at the park increases the likelihood of direct proximity at the relevant time.	Strong [E5, E6, E4] Hector is documented on the skate park sign-in sheet on the day of the murder with times matching the coroner's timeframe, and he lacks a convincing alibi for that period. His known habit of frequenting the park supports that he could plausibly be present when Roxanne was poisoned.
Means	Strong [E1, E9, E10] The homicide method was frog toxin, and a pet store owner reports Stacey recently bought an exotic poison dart frog consistent with that method. Her stated rainforest experience and interest in poison dart frogs further support capability to handle or deploy the toxin.	Moderate [E1, E2, E3] Roxanne died from poison from a rare South American frog. Hector's repeated South America travel and amphibian specialization suggest access and technical knowledge to obtain/handle toxin. However, there is no direct evidence he actually acquired a frog/toxin or possessed it locally, making means suggestive but not definitive.
Alibi	Unknown [] The record provided places Stacey at/near the scene but does not include Stacey's own account of her whereabouts during the critical window or any verifiable corroboration, so the alibi cannot be assessed beyond being unestablished here.	Weak [E6, E5] He has no convincing alibi during the relevant time window, and the sign-in sheet places him at the scene during that timeframe, undermining any unsubstantiated claim that he was elsewhere.

Table 2: M.O.M.A Analysis Results. Each dimension shows assessment level, supporting evidence IDs, and reasoning.

Evidence ID	Stacey	Hector	Diag.
E1	C	C	Low
E2	N	C	Medium
E3	N	C	Medium
E4	N	C	Low
E5	N	C	Medium
E6	N	C	Medium
E7	C	N	Medium
E8	C	N	Medium
E9	C	I	High
E10	C	N	Low
E11	C	I	High
E12	C	N	Medium
Inconsistency (I) Count	0	2	–

Table 3: Evidence-Hypothesis Consistency Matrix. C: Consistent, I: Inconsistent, N: Neutral.

Verdict	Stacey
Confidence	78
Reasoning: Both suspects had opportunity at the skate park, but Stacey is the only one tied to both a strong, case-specific motive (credible accounts of sustained design theft and confrontation) and direct means (recent purchase of an exotic poison dart frog consistent with the toxin). Hector’s travel and amphibian expertise suggest possible access and knowledge, but the record shows no comparable motive or direct evidence he acquired or possessed the poison used.	

Table 4: Judge ($\mathcal{A}_4$) Final Verdict