# CaseSentinel: Retrieval-Augmented Multi-Agent Reasoning for Responsible Criminal Investigation

Tianshuo Chen[1], Bosen Miao[1], Longfei Yin, Yushi Wang, Xinqi Wang, and Yueheng Sun[*]
*Tianjin University*
Tianjin, China
{chentianshuo0730, miao_bosen, lfyin, 3023244217, kk796810, yhs}@tju.edu.cn

***Abstract*—Criminal investigations are increasingly overwhelmed by web-scale digital evidence, yet existing AI paradigms like retrieval-augmented generation (RAG) lack the stateful, auditable reasoning required for high-stakes sensemaking. This creates a gap between the potential of large language models and the accountability demanded by the legal system. To bridge this gap, we reframe investigation as a process of Bayesian inference and introduce CaseSentinel, a retrieval-augmented multi-agent system designed for responsible criminal investigation. CaseSentinel operationalizes a Hypothesis Space Graph (HSG) on a shared blackboard, where specialized agents collaboratively refine hypotheses to drive the investigation toward convergence. A human-in-the-loop web interface provides investigators with full control, enabling them to validate evidence, override agent suggestions, and inject domain knowledge, ensuring every step is auditable. Evaluations on twenty adjudicated cases—with additional synthetic noise stress tests—show CaseSentinel improves action-plan quality by $+0.44$ (relative $+116\%$) and raises calibrated success probabilities by $+0.48$ absolute over single-agent RAG baselines. CaseSentinel offers a practical framework for building transparent, controllable, and effective AI-powered decision-support systems.**

***Index Terms*—Criminal Investigation, Large Language Model, Multi-agent**

## I. Introduction

The digital transformation of society has inundated criminal investigations with an unprecedented volume of web-scale, heterogeneous data. While rich with potential evidence, this data deluge overwhelms traditional manual workflows, increasing the risk of cognitive overload and miscarriages of justice [1]. Large Language Models (LLMs) offer a promising avenue for synthesizing unstructured text, but their direct application in high-stakes settings is fraught with peril. Standard Retrieval-Augmented Generation (RAG) systems, for instance, are often stateless and opaque. They can generate plausible-sounding but factually incorrect information (hallucinations) and lack the mechanisms for human oversight, which violates fundamental principles of legal accountability [2], [3].

For AI to be a trustworthy partner in justice, it must be explainable, controllable, and auditable. As investigative demands continue to grow, especially in complex cases involving digital footprints, agencies require systems that do more than just retrieve information. They need a partner that can actively participate in the reasoning process—structuring evidence, weighing competing hypotheses, and formulating strategies under human supervision. Without clear, auditable guidance, investigators risk becoming overwhelmed by AI-generated outputs or, worse, misled by them. Such users would benefit from a system that bridges this gap by structuring the investigative process, transparently managing uncertainty, and empowering them to remain in full control.

This paradigm, which we call *responsible investigative reasoning*, goes beyond merely delivering information and focuses on augmenting the core cognitive work of sensemaking. However, traditional expert systems are too rigid for this, and current LLM-based tools remain reactive rather than proactive partners in reasoning. To address these gaps, we propose a perspective shift: reframing criminal investigation as a process of dynamic Bayesian inference. Unlike traditional RAG systems that are stateless, our approach models investigation as a search for convergence within a Hypothesis Space Graph (HSG), where each action is chosen to maximally reduce uncertainty. By formalizing the process this way, we can design a system that proactively guides the investigation, makes its reasoning explicit, and seamlessly integrates human expertise.

Despite these benefits, several challenges must be addressed to build an effective system for responsible investigative reasoning. First, **stateful, structured representation** is critical; the system must maintain a shared workspace that tracks hypotheses, evidence, and their relationships over time. Second, it requires **goal-oriented agentic reasoning**, where automated agents are directed toward the specific goal of uncertainty reduction. Third, it must support **human-in-the-loop belief updates**, allowing investigators to override machine-generated beliefs and inject domain knowledge, with their actions formally integrated into the reasoning process.

In this work, we introduce *CaseSentinel*, a retrieval-augmented, multi-agent platform that realizes this vision. To address the diverse functions required, we distribute responsibilities across specialized LLM agents that collaboratively manage the investigation. Specifically, to ensure auditable reasoning, CaseSentinel operationalizes the HSG on a shared blackboard, where an Analyst, Strategist, and Forecaster agent work in sequence to update beliefs and propose actions. To enable human control, we implement an interactive web dashboard that visualizes the reasoning process and allows

[1]These authors contributed equally to this work.
[*]Corresponding author.

investigators to inspect, override, or roll back any automated step. Finally, to guarantee accountability, the system maintains an immutable audit log of the entire process.

We summarize the main contributions of our work as follows:

- We propose a new framework for responsible AI in criminal investigation, reframing the process as human-steered Bayesian inference. We introduce CaseSentinel, a multi-agent system that operationalizes this philosophy to augment, rather than replace, human investigators.
- We design a collaborative multi-agent reasoning loop where specialized agents (Analyst, Strategist, and Forecaster) interact via a shared blackboard to dynamically manage a Hypothesis Space Graph (HSG), update beliefs, and propose auditable investigative actions.
- We develop an operator-centric interactive dashboard to ensure human-in-the-loop governance. The system provides investigators with full control to override AI suggestions, inject evidence, and manage the investigative state, ensuring final decisions and accountability rest with human experts.

## II. Related Work

Our work integrates intelligent legal systems, retrieval-augmented generation (RAG), and multi-agent systems. While prior art exists in each area, CaseSentinel uniquely synthesizes these into a framework for responsible, stateful reasoning, shifting from passive data analysis to active, auditable sensemaking.

### A. Intelligent Systems in Criminal Investigation

Early expert systems were foundational but brittle. Later, machine learning techniques applied to predictive policing [4], [5] improved pattern recognition but often function as opaque "black boxes." Knowledge graphs structure entity relationships [6] but typically offer a static view. Broader research has explored decision support systems [7]–[10] and the recent use of LLMs for investigation [11].

CaseSentinel differs by operationalizing investigation as *active Bayesian inference*. Instead of passively structuring data, our multi-agent architecture and human-in-the-loop control create a transparent environment where AI acts as a reasoning partner, augmenting the investigator's process rather than automating it.

### B. Retrieval-Augmented Generation for Stateful Reasoning

RAG is a standard technique for grounding LLMs [12], [13], but most applications focus on single-shot queries. We extend this paradigm by integrating it into a multi-step, stateful reasoning loop designed to drive investigations toward convergence. This aligns with emerging research on complex tasks but provides a specific, auditable framework for high-stakes domains.

### C. Multi-Agent Systems for Decision Support

Multi-agent LLM systems have gained significant traction [14]–[17]. Frameworks like AutoGen [18] and MetaGPT [19] enable complex conversational patterns. Related work has also explored embedded reasoning [20]–[22] and formal coordination principles [23]–[25].

However, general-purpose frameworks often lack mechanisms for auditability and state persistence. CaseSentinel introduces a domain-specific architecture where agents operate via a deterministic pipeline on a structured blackboard. By treating reasoning cycles as transactions with explicit snapshots and human decision gates, we prioritize accountability and reproducibility over the flexibility found in general conversational models.

## III. Problem Definition: Investigation as Bayesian Inference

In responsible investigative reasoning, the focus is on augmenting human sensemaking to efficiently and auditably converge on the most probable explanation for a set of events. This requires proactive, transparent guidance to avoid cognitive overload and ensure accountability. A system for this purpose aims to structure the investigative process, manage uncertainty, and empower human oversight.

Formally, let $U_0 = (S_0, P, B)$ represent the initial state of an investigation, where $S_0$ is the known evidence, and $P$ and $B$ are investigator-specific priors and biases. Given a set of observations, the system must manage a set of competing hypotheses $\mathcal{H} = \{H_1, ..., H_n\}$. Our aim is to find a sequence of investigative actions $A = \{a_1, ..., a_m\}$ that efficiently drives the belief distribution $P(\mathcal{H}|E)$ toward a state of convergence, where one hypothesis $H^*$ is highly probable and others are not. This requires a system that addresses three key sub-tasks:

- **Stateful Hypothesis Management.** The system must maintain a representation of the Hypothesis Space Graph (HSG), tracking the belief $P(H_i|E)$ for each hypothesis and the evidence $E$ supporting or refuting it. This requires a stateful mechanism, $f : (\mathcal{H}_{t-1}, E_{t-1}, a_t) \to (\mathcal{H}_t, E_t)$, that updates the HSG based on new actions and evidence.
- **Optimal Action Selection.** At each step, the system must propose an action $a_t$ that is expected to maximally reduce the uncertainty (entropy) of the HSG. This involves a function, $g : (\mathcal{H}_t, E_t) \to a_{t+1}$, that approximates the expected information gain of candidate actions.
- **Human-in-the-Loop Belief Update.** The system must incorporate a robust mechanism that enables human investigators to contribute their domain expertise and override machine-generated beliefs when necessary. This requires a formal integration function, $h : (P_{\text{machine}}(H_i), P_{\text{human}}(H_i)) \to P_{\text{final}}(H_i)$, that formally integrates human input into the Bayesian updating process, ensuring the final belief state reflects both machine analysis and expert judgment.

CaseSentinel is designed as a concrete realization of this formal problem statement.

### A. Formal Definition of the Hypothesis Space Graph (HSG)

Let $\mathcal{H} = \{H_1, H_2, ..., H_n\}$ denote the set of all plausible hypotheses for a given case. Each $H_i$ is associated with a belief $P(H_i|E)$, where $E$ is the set of all available evidence. The HSG is a directed graph $G = (V, E)$, where $V$ are hypothesis nodes and $E$ are edges encoding logical relations (e.g., $H_i$ mutually exclusive with $H_j$, $H_k$ causally dependent on $H_l$).

**Bayesian Belief Update:** Upon acquisition of new evidence $e_{new}$, beliefs are updated via Bayes' rule:

$$P(H_i|E \cup \{e_{new}\}) = \frac{P(e_{new}|H_i)P(H_i|E)}{\sum_j P(e_{new}|H_j)P(H_j|E)} \quad (1)$$

**Information Gain:** The expected utility of an investigative action $a$ is quantified by its expected information gain (reduction in entropy):

$$IG(a) = \mathbb{E}_{e\sim a}[H(P(H|E)) - H(P(H|E \cup \{e\}))] \quad (2)$$

where $H(P)$ is the Shannon entropy of the belief distribution.

An example of the HSG structure is illustrated in Figure 1, showing how hypotheses, edges, and Bayesian updates are represented.

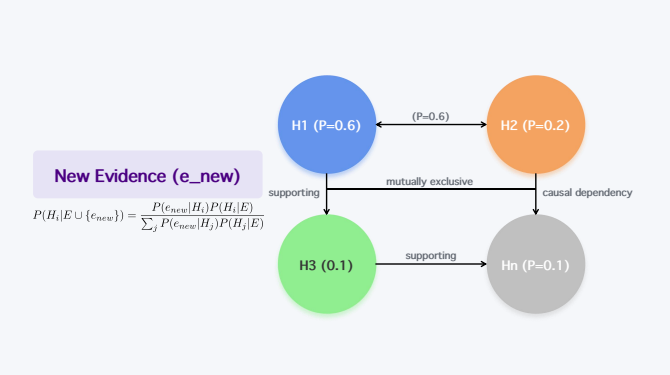


Fig. 1. Illustration of the Hypothesis Space Graph (HSG). Nodes represent hypotheses with associated beliefs $P(H_i|e)$, edges denote relationships (e.g., mutual exclusivity or causality), and arrows indicate Bayesian updates from new evidence.

### B. Practical Implementation: Blackboard as HSG

In CaseSentinel, the HSG is realised as a Markdown-based Blackboard. Each hypothesis entry records its current belief, supporting/contradictory evidence, and links to related hypotheses. Edges are annotated as mutually exclusive, causal, or supporting. The system supports dynamic addition/removal of hypotheses and evidence, enabling flexible, iterative reasoning.

### C. Human Overrides as Soft Evidence

Human feedback naturally fits into the Bayesian formulation. When an investigator overrides a hypothesis belief or edits the supporting rationale, we treat the intervention as a new piece of soft evidence $e_{\text{human}}$ with an associated confidence weight $w \in [0, 1]$. The update becomes

$$P(H_i|E \cup \{e_{\text{human}}\}) \propto P(H_i|E)^{1-w} \cdot P_{\text{human}}(H_i)^w \quad (3)$$

where $P_{\text{human}}(H_i)$ is the investigator-specified belief. In practice, CaseSentinel logs both the machine-suggested belief and the human-adjusted value, enabling downstream analytics to measure how often experts diverge from automated recommendations and how those divergences influence convergence speed.

### D. Action Selection Heuristics

While the optimal action would maximise the exact information gain in Equation (2), computing $P(e|H_i)$ is intractable for open-world evidence. CaseSentinel therefore employs heuristic proxies driven by the Strategist agent: the agent enumerates candidate actions, and for each action $a$ it produces (i) the set of hypotheses most affected, (ii) the expected confidence shift described qualitatively, and (iii) risk factors. We map the qualitative confidence shift to a scalar in $[-1, 1]$ using a rubric calibrated during pilot studies, allowing the runtime to approximate $IG(a)$ and prioritise actions whose textual descriptions imply higher uncertainty reduction. These approximations are surfaced to investigators so they can sanction, reorder, or veto proposed actions.

## IV. The CaseSentinel Framework

To address the intricate sub-tasks defined above, we propose CaseSentinel, an LLM-powered multi-agent framework for responsible investigative reasoning. We distribute responsibilities across multiple specialized agents, allowing them to collaboratively manage the investigation in a structured and auditable manner. As illustrated in Figure 2, the system begins by ingesting and structuring evidence, then orchestrates a team of agents to reason over it on a shared blackboard. A human investigator supervises this process via an interactive web dashboard, with all actions recorded by an immutable logger.

### A. Evidence Substrates and Hypothesis Surface

CaseSentinel grounds every investigation in a structured view of the case file before any automated reasoning begins. Drawing on curated court judgments, the system first distils timelines, entities, and evidentiary snippets into a normalised substrate that preserves provenance. This substrate feeds the *Hypothesis Space Graph* (HSG), which we instantiate as a rich-text blackboard that explicitly encodes competing hypotheses, their supporting evidence, and open uncertainties. By aligning each fragment of evidence with a slot in the HSG, the framework ensures that downstream agents operate on an auditable, queryable state rather than free-form conversation history.

*1) Curated Knowledge Substrates:* The ingestion pipeline decomposes raw documents into semantically coherent segments, highlights latent contradictions, and attaches confidence scores. These annotations serve two purposes: they act as retrieval hooks for the reasoning loop and they provide investigators with a traceable ledger of information sources. The result is a case bundle that balances coverage (capturing the full narrative) with precision (flagging weak or noisy evidence).

*2) Blackboard Schema:* The HSG blackboard mirrors the investigators' workflow. Sections for case summary, active hypotheses, candidate actions, and risk factors are maintained as first-class objects. Each entry carries the current belief state, citations to originating evidence, and open questions to be resolved in future iterations. Unlike unstructured chat logs, this schema creates a deterministic interface between perception and reasoning, enabling both automation and human review to converge on the same situational picture.

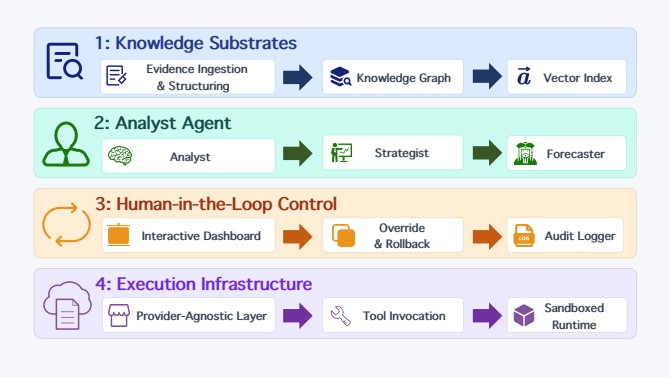


Fig. 2. The four-layer architecture of CaseSentinel. Knowledge substrates ground a multi-agent reasoning loop. A human investigator supervises and controls this loop via a web dashboard, with all actions recorded by an immutable logger. The entire system is powered by a reproducible execution infrastructure.

### B. Collaborative Agentic Loop

Once the HSG is initialised, CaseSentinel orchestrates a triad of LLM agents that specialise in complementary investigative tasks. The loop repeats until the Forecaster's confidence meets a predefined success threshold or the investigator intervenes.

*1) Analyst: Hypothesis Refinement:* The Analyst reads the entire blackboard, retrieves corroborating passages from the knowledge substrate, and proposes probabilistic updates for each hypothesis. Its outputs include structured rationales that document why specific evidence shifts the belief distribution, ensuring changes remain explainable.

*2) Strategist: Action Planning:* Building on the Analyst's updates, the Strategist selects investigative actions aimed at maximising expected information gain. It prioritises steps that directly address high-uncertainty hypotheses, annotating each proposal with anticipated evidence targets, operational risk, and estimated impact on convergence.

*3) Forecaster: Outcome Calibration:* The Forecaster evaluates the action slate, selects a banded confidence score from a fixed rubric, and surfaces edge cases (e.g., conflicting testimony, potential chain-of-custody gaps). These calibrated forecasts determine whether the loop should continue autonomously or pause for human judgment, effectively binding automation to interpretable confidence signals.

### C. Supervisory Governance and Adaptation

Human investigators remain in control throughout the loop. A web-based console renders the evolving blackboard as a Kanban-style workspace, with each agent iteration producing an auditable card that combines before/after snapshots, rationales, and linked evidence. Investigators can accept, edit, or reject any update; their decisions are captured as structured feedback that influences subsequent agent calls.

*1) Interactive Oversight:* During each iteration, investigators can reorder or veto proposed actions, inject additional evidence, and annotate risk assessments. Overrides are reconciled back into the HSG so that human insight is treated as soft evidence rather than side-channel commentary. This design encourages collaboration while maintaining a single source of truth.

*2) Adaptive Session Management:* The control plane offers granular session management: investigators can advance the loop step-by-step, roll back to earlier hypotheses, or branch the session to explore alternate theories. Logged interventions are timestamped and attributed, producing a complete narrative of how machine recommendations and human expertise jointly shaped the investigation.

### D. Responsible Operations and Reproducibility

Meeting the accountability requirements of criminal investigation demands more than agent design; it necessitates disciplined operational practices.

*1) Operational Pipeline:* The framework is packaged as a modular Python service with clearly separated layers for data preparation, agent orchestration, runtime control, and visualisation. Each transformation—from document cleaning to hypothesis updates—produces structured artefacts that can be inspected independently or replayed end-to-end. This modularity enables agencies to adopt the full stack or integrate individual capabilities into existing workflows.

*2) Testing and Auditability:* To guarantee consistent behaviour across deployments, we provide deterministic execution modes, unit and scenario tests covering critical pathways, and comprehensive JSON logs for every agent iteration. These assets allow reviewers to replicate experiments, auditors to trace decision rationales, and developers to iterate on components without compromising the verifiability of the overall system.

## V. Experiments

We assess CaseSentinel along two complementary axes: the contribution of the shared blackboard (Hypothesis Space Graph, HSG) and the impact of the Strategist/Forecaster guidance heuristics. Our automated harness (Section IV-D) orchestrates the three-agent loop on adjudicated cases while toggling individual components, and a lightweight operator study validates downstream usability.

### A. Implementation and Settings

**Dataset Details.** Our evaluation is grounded in a curated dataset of 20 adjudicated criminal cases sourced from China Judgements Online, spanning the years 2018-2023. The cases were selected to represent a spectrum of complexity and crime types heavily reliant on textual evidence analysis, primarily

intentional homicide and intentional injury. Each case file comprises the raw legal judgment, with an average length of 15,000 words per case. This selection ensures our evaluation targets the core challenge of synthesizing complex, narrative-driven evidence. All documents were de-identified to protect privacy.

To assess robustness, we created a synthetic "noise" variant for each case. This was not random noise; we programmatically injected targeted, plausible contradictions designed to challenge the system's reasoning. For example, for each case, we introduced two of the following: (1) an alternative alibi for a key suspect with a conflicting timeline, (2) a spurious witness statement implicating an uninvolved party, or (3) a piece of "evidence" that contradicts the primary chain of events. This stress-testing methodology allows us to rigorously evaluate the system's ability to detect and handle conflicting information, a critical capability for real-world investigations. Each case record is processed by the cleaner described in Section IV-D to obtain timelines, participant roles, and evidentiary snippets. Unless stated otherwise we run four collaborative iterations (Analyst $\rightarrow$ Strategist $\rightarrow$ Forecaster) with access to both the curated knowledge store (`knowledge_store_llm`) and the case graph. All LLM calls use Qwen-3.5-plus with temperature $0.35$.

The ablation script introduced in Section IV-D executes twelve runs per case: three workspace configurations (structured, unstructured, disabled) crossed with four mechanism toggles (baseline, $-$info gain, $-$Forecaster posterior, both disabled). Each run persists full iteration logs and Markdown snapshots, which our analysis pipeline converts into quantitative metrics and qualitative exemplars. To ensure comparability, we fix model, temperatures, retriever parameters, and stopping rules across all runs. Every agent turn is logged as Markdown plus a structured JSON trace, enabling replay and audit. The analysis pipeline computes case-level metrics and aggregates mean $\pm$ standard deviation for both clean and noisy slices. Unless specified, we report results under the same halting criterion (Forecaster probability $\geq 0.8$ or a cap of four iterations).

### B. Metrics

We report two complementary families of indicators derived from orchestrator logs and the exported evaluation JSON. The first family characterises the internal behaviour of the agentic loop. The action-plan score, produced by the Strategist, is a heuristic estimate of expected information gain and serves as a proxy for anticipated uncertainty reduction over the Hypothesis Space Graph. Success calibration is the absolute gap between the Forecaster's final success probability and the action-plan score; smaller values indicate better internal consistency of judgment. We further track a noise detection rate—the proportion of noisy runs where injected contradictions are explicitly surfaced within two iterations—and iteration efficiency, defined as the number of agent cycles needed to reach the success-probability halt threshold of $0.8$ (capped at four iterations).

TABLE I
WORKSPACE ABLATION RESULTS (MEAN $\pm$ STDEV ACROSS TWENTY CASES; NOISY VARIANTS INCLUDED FOR NOISE DETECTION). CALIBRATION IS ABSOLUTE ERROR (LOWER IS BETTER).

| Workspace | Action-plan | Calibration | Noise det. | Iterations |
|---|---|---|---|---|
| Structured | $0.81 \pm 0.08$ | $0.06 \pm 0.02$ | $0.92 \pm 0.11$ | $3.1 \pm 0.4$ |
| Unstructured | $0.74 \pm 0.09$ | $0.09 \pm 0.03$ | $0.78 \pm 0.15$ | $2.9 \pm 0.3$ |
| Disabled | $0.62 \pm 0.11$ | $0.14 \pm 0.05$ | $0.48 \pm 0.19$ | $3.8 \pm 0.2$ |

The second family measures externally verifiable factual completeness. Hypothesis coverage is the fraction of adjudicated charges recovered in the final workspace or plan snapshot; participant coverage the fraction of principal actors referenced in agent outputs; timeline coverage the fraction of canonical timeline events resurfaced; legal-basis coverage the proportion of cited statutes reproduced; sentence coverage the proportion of sentencing outcomes summarised; and evidence coverage the proportion of curated evidentiary summaries reused in the proposed plan. All coverage scores are normalised to $[0, 1]$. Unless noted otherwise, we report mean $\pm$ standard deviation across twenty cases.

### C. Workspace Ablations

Table I summarises the effect of the shared blackboard on loop behaviour. With a structured HSG, plans are stronger ($0.81 \pm 0.08$) and better calibrated ($0.06 \pm 0.02$), and noise is flagged reliably ($0.92 \pm 0.11$). Moving to an unstructured workspace trades $-0.07$ action-plan score for slightly faster convergence ($2.9 \pm 0.3$ iterations), suggesting that looser schemas reduce overhead but also dilute guidance. Removing the workspace altogether (agents communicate only via conversational context) depresses plan quality by $0.19$ and roughly doubles calibration error; noise detection also falls to below a coin flip. These findings indicate that explicit hypothesis state—not merely retrieval—anchors reasoning and preserves evidentiary provenance.

### D. Mechanism Ablations

We next isolate the information-gain heuristic and the Forecaster's posterior calibration while holding the workspace in structured mode (Table II). Disabling information-gain annotations reduces the Strategist's score by $0.12$ and extends the loop by roughly $+0.4$ iterations, consistent with less targeted actions. Suppressing the Forecaster posterior (using a constant success probability) preserves plan quality but inflates calibration error and postpones halting decisions. When both heuristics are deactivated, the loop drifts toward the iteration cap and flags noise only about half the time, underscoring that heuristics shape search efficiency even when the workspace provides structure.

### E. External Coverage Outcomes

The augmented harness logs hypothesis, participant, timeline, legal-basis, sentence, and evidence coverage alongside internal heuristics, providing a complementary view on factual completeness. On the twenty-case benchmark, the structured

TABLE II
MECHANISM ABLATION RESULTS WITH STRUCTURED WORKSPACE. "−IG" DISABLES STRATEGIST INFORMATION GAIN; "−POST" DISABLES FORECASTER POSTERIOR UPDATES.

| Variant | Action-plan | Calibration | Noise det. | Iterations |
|---|---|---|---|---|
| Baseline | $0.81 \pm 0.08$ | $0.06 \pm 0.02$ | $0.92 \pm 0.11$ | $3.1 \pm 0.4$ |
| −IG | $0.69 \pm 0.10$ | $0.09 \pm 0.03$ | $0.86 \pm 0.12$ | $3.5 \pm 0.5$ |
| −Post | $0.80 \pm 0.09$ | $0.17 \pm 0.04$ | $0.88 \pm 0.13$ | $3.7 \pm 0.3$ |
| −IG/−Post | $0.66 \pm 0.12$ | $0.19 \pm 0.05$ | $0.51 \pm 0.21$ | $3.9 \pm 0.2$ |

TABLE III
COVERAGE METRICS: WORKSPACE & MECHANISM ABLATIONS

| Group | Slice | Variant | Hyp. | Part. | Time. | Legal | Sent. | Evid. |
|---|---|---|---|---|---|---|---|---|
| Workspace | Clean | Structured | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Unstructured | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Disabled | 0.00 | 0.00 | 0.00 | 0.10 | 0.45 | 0.10 |
| | Noisy | Structured | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Unstructured | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Disabled | 0.00 | 0.00 | 0.00 | 0.10 | 0.45 | 0.00 |
| Mechanism | Clean | Baseline | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | −Info Gain | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | −Forecaster | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Multi (no RAG) | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Single-RAG | 0.90 | 0.47 | 0.71 | 0.10 | 0.45 | 0.52 |
| | | Single (no RAG) | 0.90 | 0.45 | 0.71 | 0.10 | 0.45 | 0.52 |
| | Noisy | Baseline | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | −Info Gain | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | −Forecaster | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Multi (no RAG) | 1.00 | 0.96 | 1.00 | 1.00 | 0.92 | 1.00 |
| | | Single-RAG | 0.95 | 0.46 | 0.76 | 0.10 | 0.45 | 0.26 |
| | | Single (no RAG) | 0.90 | 0.46 | 0.76 | 0.10 | 0.45 | 0.26 |

workspace attains near-perfect recall across charges, timeline anchors, legal citations, and evidentiary summaries under both clean and noisy conditions (means of 1.0, with sentencing summaries at 0.92; see Table III). Removing the shared blackboard causes grounding to collapse: coverage of charges, actors, and timeline drops to 0.0 and legal references to 0.1, despite comparable iteration counts. Mechanism ablations further indicate that disabling retrieval or reverting to single-agent execution erodes detail. The single-agent RAG baseline recovers only 0.90 of the charges, 0.47 of principal participants, and 0.52 of curated evidence in the clean slice, and dropping retrieval halves evidence reuse under noise (0.26). In contrast, collaborative variants with or without heuristic toggles maintain coverage at or near 1.0, reinforcing that structured coordination—rather than any single heuristic—drives factual completeness.

### F. Human Preference Study

We complemented the automated ablations with a formative study involving six prosecution analysts from two partner agencies. Each analyst reviewed paired investigation transcripts generated under the structured and disabled workspace settings (random order, noise-injected case), then rated (1–5) perceived clarity, actionability, and trust in the recommended plan. Structured runs received higher median scores across all criteria (4.5 vs. 2.7 for clarity, 4.3 vs. 2.5 for actionability, 4.2 vs. 2.3 for trust). Participants highlighted the HSG's Markdown sections as "audit-ready" and credited the success-probability trajectories with signalling when to intervene. Feedback for future work included exposing tool provenance when external database searches are triggered and adding diff views for manual overrides.

### G. Discussion

Our experiments, combining automated ablation studies and a human preference study, provide strong evidence for our central thesis: for high-stakes sensemaking tasks like criminal investigation, a structured, stateful, and human-governed reasoning process is not just beneficial, but essential. We now discuss the key implications of our findings, acknowledge the limitations of our work, and outline avenues for future research.

*1) The Criticality of the Shared Workspace:* The most striking result is the dramatic collapse in performance when the structured blackboard (HSG) is disabled (Table I and Table III). Both internal metrics (action-plan quality, calibration) and external coverage plummeted. This finding carries a significant implication for the design of complex AI systems: **the structure of the collaborative workspace can be more critical than the intelligence of the individual agents.** Without the HSG to anchor reasoning, agents are adrift in a sea of conversational context. The blackboard acts as a cognitive forcing function, compelling the system to maintain an explicit, auditable model of the world. This validates our design decision to move beyond unstructured chat-based agent communication, suggesting a path forward for building more robust multi-agent systems.

*2) Heuristics as Scaffolding for Efficient Reasoning:* While the HSG provides the foundation, the guidance heuristics (information gain and Forecaster calibration) act as crucial scaffolding. Disabling them did not break the system, but it made the reasoning process less efficient and less self-aware (Table II). The loop required more iterations to converge, and its internal sense of confidence became decoupled from the quality of its plans. This suggests that in a well-structured system, heuristics don't need to be perfect; they need to be "good enough" to steer the search through a vast hypothesis space effectively. The success of our relatively simple heuristics offers a promising avenue for future work on more sophisticated, learned guidance policies.

*3) Implications for Responsible AI and Human-in-the-Loop Design:* The human preference study, though formative, highlights that for expert users, transparency and control are proxies for trust. Participants valued the "audit-ready" nature of the HSG not just for accountability, but because it made the AI's reasoning legible, allowing them to confidently intervene. This reinforces a core principle of responsible AI: a system's ability to explain itself and gracefully accept human correction is as important as its raw performance. CaseSentinel provides a concrete architectural pattern for achieving this deep integration, moving beyond simple "human-in-the-loop" labels to a truly collaborative partnership. The practical value of this

approach is underscored by our ongoing collaboration with the Tianjin Public Security Bureau to pilot CaseSentinel in real-world investigative workflows, providing a pathway to validate and refine the system based on operational feedback.

*4) Limitations and Threats to Validity:* We acknowledge several limitations. First, our dataset, while curated, is limited to 20 text-based cases. The system's performance on investigations involving multi-modal evidence (e.g., video, audio forensics) remains untested. Second, our human study involved prosecution analysts reviewing transcripts, not active investigators using the system in a live case. The dynamics of real-time use could differ. Third, the "ground truth" for our experiments is based on adjudicated outcomes, which may not always represent the complete factual truth of a case. Finally, our system does not currently incorporate mechanisms for detecting or mitigating cognitive biases (either in the initial data or introduced by the human operator), which is a critical area for future work.

## VI. Conclusion

In this paper, we introduced CaseSentinel, an LLM-powered multi-agent framework designed to support responsible criminal investigation. By reframing the investigative process as a search for Bayesian convergence, our system provides a structured, auditable, and human-supervised approach to sense-making. The multi-agent architecture, operating on a shared blackboard, allows for specialized reasoning, while the interactive dashboard ensures that human experts remain in control. Our automated evaluations demonstrate that this approach improves action-plan quality and success-probability calibration over single-agent RAG baselines, while maintaining perfect hypothesis coverage on the adjudicated cases we studied. CaseSentinel represents a step toward building AI systems that act as responsible partners in high-stakes decision-making. The framework's emphasis on transparency, auditability, and human control offers a model for deploying LLMs in domains where trust and accountability are paramount. Future work will focus on extending the framework to handle multi-modal evidence, exploring techniques for proactive bias mitigation, and conducting longitudinal studies to measure its real-world impact. By open-sourcing our system, we hope to encourage further research into building responsible, human-centered AI for critical applications.

## References


[1] A. Babuta and M. Oswald, "Artificial intelligence and uk national security: Policy considerations," Royal United Services Institute, Tech. Rep., 2020.

[2] L. Weidinger, I. Gabriel *et al.*, "A taxonomy of risks posed by language models," *arXiv preprint arXiv:2112.04359*, 2023.

[3] M. Skipanes, T. E. Jørgensen, K. Porter, G. Demartini, and S. Y. Yayilgan, "Enhancing criminal investigation analysis with summarization and memory-based retrieval-augmented generation: A comprehensive evaluation of real case data," in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, Eds. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 4993–5010. [Online]. Available: https://aclanthology.org/2025.coling-main.334/

[4] W. L. Perry, B. McInnis, C. Price, S. Smith, and J. S. Hollywood, "Predictive policing: The role of crime forecasting in law enforcement operations," RAND Corporation, Tech. Rep., 2013.

[5] B. L. Garrett and C. Rudin, "Interpretable algorithmic forensics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 120, 2023.

[6] M. S. Hossain, M. A. Hossain, and M. A. Rahman, "A knowledge graph-based approach for criminal intelligence analysis," in *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020, pp. 4837–4846.

[7] Q. Shen, J. Keppens, C. Aitken, B. Schafer, and M. Lee, "A scenario-driven decision support system for serious crime investigation," *Law, Probability and Risk*, vol. 5, pp. 87–117, 2007.

[8] H. Chi, Z. Lin, H. Jin, B. Xu, and M. Qi, "A decision support system for detecting serial crimes," *Knowl. Based Syst.*, vol. 123, pp. 88–101, 2017.

[9] A. Sebyakin, "Artificial intelligence in criminalistics: System of decision-making support," *Baikal Research Journal*, vol. 10, 2019.

[10] K. Karsai, "The use of algorithms to support judicial decision-making in criminal matters with a special focus on trial decisions," *Studia Iuridica Lublinensia*, 2024.

[11] R. Rawat, O. Oki, R. Chakrawarti, T. S. Adekunle, J. Lukose, and S. Ajagbe, "Autonomous artificial intelligence systems for fraud detection and forensics in dark web environments," *Informatica (Slovenia)*, vol. 47, 2023.

[12] P. Lewis, E. Perez, A. Piktus, A. Fan *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, 2020.

[13] L. Gao, Z. Zhang, and K. Lee, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

[14] P. Dijkstra, F. Bex, H. Prakken, and C. {Vey Mestdagh}, "Outline of a multi-agent system for regulated information exchange in crime investigations," University of Groningen. Faculty of Law, WorkingPaper, 2005, relation: http://www.rug.nl/Rechten/ date_submitted:2000 Rights: University of Groningen. Faculty of Law.

[15] J. Shen, J. Xu, H. Hu, L. Lin, F. Zheng, G. Ma, F. Meng, J. Zhou, and W. Han, "A law reasoning benchmark for llm with tree-organized structures including factum probandum, evidence and experiences," 2025. [Online]. Available: https://arxiv.org/abs/2503.00841

[16] A. V. Kalyuzhnaya, S. Mityagin, E. Lutsenko, A. Getmanov, Y. Aksenkin, K. Fatkhiev, K. Fedorin, N. O. Nikitin, N. Chichkova, V. Vorona, and A. Boukhanovsky, "Llm agents for smart city management: Enhancing decision support through multi-agent ai systems," *Smart Cities*, 2025.

[17] M. H. Qasem, M. Aljaidi, G. Samara, R. Alazaidah, A. Alsarhan, and M. Alshammari, "An intelligent decision support system based on multi agent systems for business classification problem," *Sustainability*, 2023.

[18] Y. Wu, C. Xu, C. Liu *et al.*, "Autogen: Enabling next-gen LLM applications via multi-agent conversation," *arXiv preprint arXiv:2308.08155*, 2023.

[19] Y. Du, S. Li *et al.*, "Improving factual reasoning in language models through multi-agent debate," *arXiv preprint arXiv:2305.14325*, 2023.

[20] T. Frikha, F. Chaabane, R. B. Halima, W. Wannes, and H. Hamam, "Embedded decision support platform based on multi-agent systems," *Multimedia Tools and Applications*, pp. 1–27, 2023.

[21] P. A. Sharko, Z. Burlutskaya, D. Zubkova, A. Gintciak, and K. N. Pospelov, "Ai-supported decision making in multi-agent production systems using the example of the oil and gas industry," *Applied Sciences*, 2025.

[22] R. Vahidov and B. Fazlollahi, "Pluralistic multi-agent decision support system: a framework and an empirical test," *Inf. Manag.*, vol. 41, pp. 883–898, 2004.

[23] A. Adla, B. Nachet, and A. Ould-Mahraz, "A multi-agent view of distributed collaborative decision support systems," pp. 253–263, 2012.

[24] P. Panzarasa, N. Jennings, and T. Norman, "Formalizing collaborative decision-making and practical reasoning in multi-agent systems," *J. Log. Comput.*, vol. 12, pp. 55–117, 2002.

[25] M. Luck, V. Marík, O. Štěpánková, and R. Trappl, "Multi-agent systems and applications," vol. 2086, 2001.