MAGR-FI: LLM-Based Multi-Agent Game-Theoretic Reasoning Framework for Fire Investigation

Tao Chen
Tsinghua University
chentao.b@tsinghua.edu.cn

Yuhao Zhang
Hefei Institute for Public Safety Research
Tsinghua University
zhangyuhao@tsinghua-hf.edu.cn

Lida Huang
Tsinghua University
huanglida@tsinghua.edu.cn

Jiakun Dai
Hefei Institute for Public Safety Research
Tsinghua University
daijiakun@tsinghua-hf.edu.cn

Feihu Sun
Hefei Institute for Public Safety Research, Tsinghua University
sunfeihu@tsinghua-hf.edu.cn

ABSTRACT

Fire investigation relies heavily on expert experience, rendering it susceptible to cognitive biases and inefficient in processing heterogeneous information. This study proposes an LLM-Based Multi-Agent Game-Theoretic Reasoning Framework for Fire Investigation (MAGR-FI), which integrates the Fire Triangle model from combustion science with the Man-Machine-Environment-Management (MMEM) framework from system safety theory to establish a structured mechanism for hypothesis generation and adversarial validation. The framework employs three specialized agents—Proponent, Skeptic, and Arbiter—engaged in a two-phase dynamic game, enabling interpretable reasoning from unstructured investigative texts to high-confidence causal conclusions. Evaluated on a test set of 1,051 real-world cases, MAGR-FI significantly outperforms baseline large language models, improving average scores from 6.95 to 8.51 and achieving a 120% performance gain on complex cases, thereby providing a reliable, transparent, and auditable intelligent decision-support tool for fire investigation and transforming retrospective causal analysis into actionable knowledge assets for crisis learning and proactive prevention.

Keywords

Fire investigation, Causal reasoning, Multi-agent systems, Game-theoretic reasoning, LLM

INTRODUCTION

Fire incidents, as one of the most frequent and destructive categories of public safety emergencies, remain a priority focus in global emergency management. Accurate and timely causal attribution not only underpins legal accountability but also serves as a critical feedback mechanism for organizational crisis learning—informing building code revisions, emergency protocol updates, and professional training enhancement. This conversion of retrospective analysis into prospective risk prevention closes the loop between post-disaster recovery and future preparedness, a core tenet of the ISCRAM crisis lifecycle perspective.

Yet fire investigation faces severe methodological challenges. Fire scenes exhibit high complexity and uncertainty due to the destructive effects of fire on physical evidence, heterogeneous information sources, and multi-factor coupled causation chains (Liu et al., 2022; Zhu et al., 2025). Traditional investigation relies heavily on experts'

tacit knowledge and intuition (Okoli et al., 2016), but this "human–knowledge" model suffers from three structural limitations:

- ***The knowledge inheritance predicament****: expert experience resists formalization, causing high inter-investigator variability and slow, costly training of novices.*
- ***The cognitive bias predicament****: confirmation bias and anchoring often lead experts to fixate on a single explanation, selectively seeking supporting evidence while ignoring alternatives—risking attribution errors or miscarriages of justice.*
- ***The information processing predicament****: investigations draw from heterogeneous sources—objective records (scene notes, forensic reports) and subjective accounts (statements, testimonies)—the latter often distorted by memory gaps, bias, or vagueness, undermining reliability.*

Recent AI advances offer pathways for intelligent transformation. Early efforts used rule-based expert systems with fixed "if–then" logic (Iliadis et al., 2002), but rigid rules fail to capture the context-dependent uncertainty of real fire causation. Later, machine learning enabled trace and image analysis (Sankarasubramanian & Ganesh, 2020), yet these "black-box" models lack interpretability, falling short of legal and scientific scrutiny standards.

More recently, Large Language Models (LLMs) show promise in semantic understanding and preliminary reasoning from unstructured texts (Ingle & Gab-Kim, 2025). However, current approaches still struggle to meet the rigor, robustness, and accountability demands of real-world fire investigation.

Against this backdrop, this study proposes a fire investigation reasoning framework based on multi-agent game theory. The research addresses two core questions:

How can unstructured investigation texts be mapped to structured information amenable to logical reasoning, while integrating theoretical foundations from fire science?How can adversarial game mechanisms among multiple agents enable dynamic evaluation of competing fire origin hypotheses and reinforcement of the most robust explanatory pathway?

The resulting agent-based architecture functions as a credible, interpretable, and traceable auxiliary reasoning tool, holding the potential to enhance attribution accuracy and analytical depth while advancing fire investigation from an experience-driven practice toward human–machine collaboration and data-augmented intelligence.

BACKGROUND

Fire Investigation and Cause Analysis

Existing fire investigation research splits into two branches. Engineering practice focuses on optimizing hardware and methods—such as scene inspection, forensic identification, and fire trace analysis—yielding mature technical standards for on-site investigation (Hine, 2004; Lentini, 2018; Munday & Gardiner, 2013). Theoretical modeling explores causation patterns via historical case statistics, fault/event tree logic, and knowledge graph–based element extraction (Goh & Chua, 2010; Xu, 2024). Systems like Case-Based Reasoning (Liu, 2009) and ontology-driven frameworks (Chandra et al., 2025) support case retrieval and knowledge reuse.

However, these approaches largely perform static information extraction or pattern matching, showing severe instability with contradictory testimonies or missing evidence. Recent work using graph neural networks for causation discovery (Zhao et al., 2024) remains limited to static structural learning, lacking dynamic validation mechanisms.

AI and Multi-Agent Systems in Crisis Management

Deep learning, knowledge graphs, and computer vision have enhanced emergency management—enabling real-time monitoring, situational awareness, and resource allocation. In fire investigation specifically, digital methods advance data processing and precedent retrieval: Mirończuk et al. (2020) used NLP to extract ignition elements from reports; Yan et al. (2021) built ontology-based knowledge graphs for case retrieval.

Recently, large language models and intelligent agents have broadened crisis informatics applications. Multi-agent systems, leveraging distributed decision-making, enable decentralized problem-solving across domains: adversarial debate in legal argumentation (Zhang & Ashley, 2025), multi-modal data fusion in healthcare (Nimbalkar et al., 2025), and crisis communication modeling within ISCRAM (Chahine et al., 2022). Yet, multi-agent game-theoretic reasoning for fire incident attribution—particularly through physically grounded, structured adversarial debate—remains unexplored.

The Fire Triangle and MMEM Framework

The Fire Triangle—fuel, oxidizer, and ignition source—is the foundational model of combustion science, stating that all three must coexist and interact to sustain fire. These constitute the physical necessary conditions: fuel determines fire load, oxidizer affects burn rate, and the ignition source supplies initial energy. In investigation, any credible hypothesis must explain how these elements converged spatiotemporally; omission of any renders it physically invalid (Drysdale, 2011).

However, the Fire Triangle model can only explain "how fire burns," not "why the accident occurred at this particular time and place"—that is, it cannot explain why system defense mechanisms failed to prevent the coupling of the three elements. Real fires stem from multiple layers of protection failure. The Man-Machine-Environment-Management (MMEM) theory complements this by attributing accidents to interactions among personnel, equipment, environment, and management—whose defects form the organizational sufficient conditions for incidents (Reason, 1990).

Table 1. MMEM Theoretical Framework and Its Role in Hypothesis Generation

Dimension	Core Question	Role in Hypothesis Generation
Man	Were operation, monitoring, and emergency behaviors compliant?	Explains "why not detected/handled in time" (e.g., asleep, misjudgment)
Machine	Was equipment reliable? Were protective devices effective?	Explains "why energy source became uncontrolled" (e.g., circuit breaker failure, line aging)
Environment	Did spatial layout, ventilation, or climate exacerbate risk?	Explains "why fire spread rapidly" (e.g., enclosed space, strong wind)
Management	Were systems, training, and maintenance adequate?	Explains "systemic vulnerabilities" (e.g., failure to replace aging outlets, lack of fire drills)

The MMEM framework does not directly cause fires but determines whether the Fire Triangle's three elements can couple in a given time and space. Thus, hypothesis generation must examine Man-Machine-Environment-Management conditions to strengthen explanatory depth. For instance, an "electrical short circuit" carries entirely different causal and liability implications under regular maintenance versus ten years of neglect.

In summary, the Fire Triangle ensures physical plausibility (necessary conditions), while MMEM ensures systemic completeness (sufficient organizational conditions). The two are complementary and synergistic, jointly forming the complete theoretical foundation for fire causation hypothesis generation.

PROPOSED FRAMEWORK

Framework Overview

Addressing the core challenges identified in fire accident investigation, this study proposes an **LLM-based Multi-Agent Game-Theoretic Reasoning Framework for Fire Investigation(MAGR-FI)**. Grounded in combustion science and system safety theory, this framework decouples the traditional investigation process—highly dependent on individual expert experience—into four interlocking, logically progressive functional modules, achieving fully computable, interpretable, and bias-resistant reasoning from unstructured raw investigation materials to standardized, high-confidence causation conclusions. The overall technical workflow and inter-module coordination are illustrated in Figure 1.

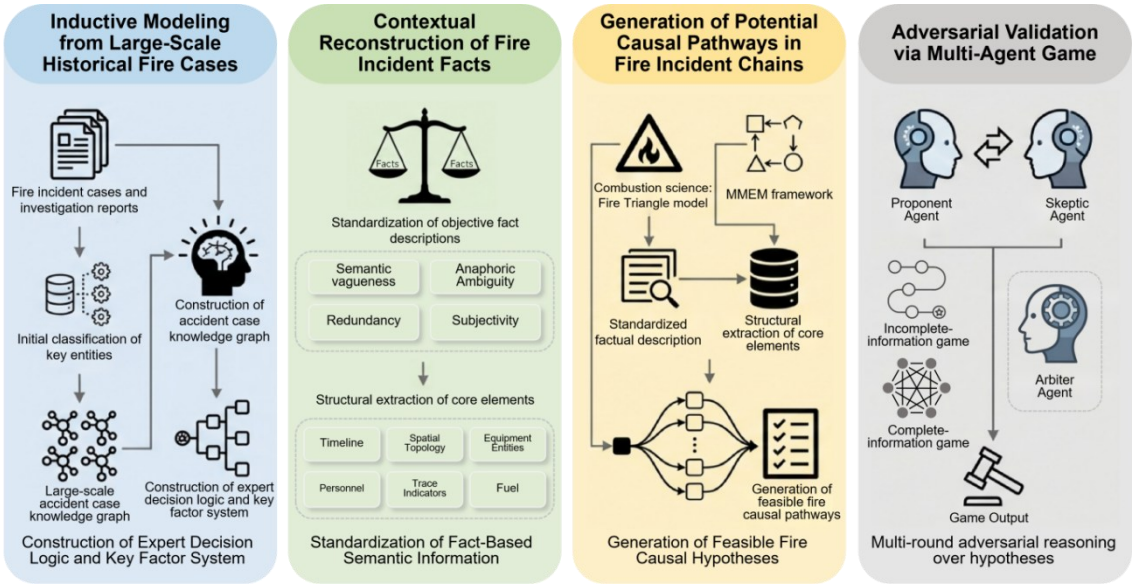

Figure 1. Overall technical flow of the proposed LLM-Based Multi-Agent Game-Theoretic Reasoning Framework for Fire Investigation (MAGR-FI)

As shown in Figure 1, the framework forms a complete reasoning pipeline from top to bottom: First, a core element system for fire investigation is constructed based on large-scale historical cases and expert consensus, establishing the domain knowledge foundation for the entire reasoning process. Second, context information modeling accomplishes denoising, standardization, and structured transformation of multi-source heterogeneous raw investigation data, forming an unbiased factual basis for reasoning. Third, through the Fire Triangle-MMEM cross matrix, systematically exhaustive generation of feasible causation pathways is realized, constructing a logically coherent initial hypothesis set. Finally, via a two-phase dynamic multi-agent game mechanism, adversarial validation and filtering of competing hypotheses are completed, outputting high-confidence causation conclusions together with comprehensive reasoning archives. The following sections elaborate on the design logic, implementation mechanisms, and core functionalities of each framework module in detail.

Step 1: Core Element System Construction

The selection of structured elements is not derived from a priori theoretical deduction but through systematic induction of common factors across a large corpus of historical fire incident reports. This ensures that the resulting element system is empirically grounded and practically oriented, specifically tailored to support causal attribution in fire investigations.

This study collected 1,263 de-identified fire investigation reports spanning 2016 to 2025, covering major fire types: building fires (55.6%), industrial fires (31.2%), and other categories (13.2%). We applied GraphRAG to construct a domain-wide knowledge graph over this corpus. A domain-finetuned LLM performed multi-round, prompt-based semantic parsing to analyze each sentence, automatically identifying key entities and assigning them to candidate element categories. Initial entity clusters—spanning over 20 candidate dimensions such as personnel, physical evidence, and fire patterns—were then refined through semantic similarity clustering.

Following manual review, redundancy elimination, and expert-guided consolidation, the system was distilled into six core dimensions:

- Timeline (32.7% of all entity nodes),
- Spatial Topology (28.4%),
- Equipment Entities (39.2%),
- Combustibles (20.9%),
- Personnel (22.1%),
- Physical Evidence of Fire Pattern (19.5%).

This six-dimensional framework achieves a coverage rate of 95.7% across nearly 2 million extracted entity nodes, demonstrating strong empirical completeness and generalizability.

To quantify each dimension's contribution to inference confidence, we conducted a counterfactual ablation experiment: for 100 high-confidence cases, we systematically removed one dimension at a time while preserving all others, then assessed the impact on conclusion reliability. Results revealed critical discriminative power:

- Removing Physical Evidence of Fire Pattern caused 83% of high-confidence cases to degrade to "multiple plausible causes" or "insufficient evidence";
- Omitting equipment entities prevented 97% of cases from identifying a specific ignition device;
- Excluding timeline information led to unresolved temporal contradictions in 53% of cases (e.g., equipment failure reported after fire suppression).

These findings confirm that each dimension plays a distinct role in disambiguating competing hypotheses and enhancing attribution uniqueness.

Notably, our analysis also revealed that information quality strongly affects conclusion reliability. Objective records (scene notes, reports) are reproducible and verifiable; subjective accounts (statements, testimonies) often contain vague language ("possibly," "seemed") and contradictions. Cases with >50% subjective content (n = 41) triggered an average of 4.8 additional technical measures (e.g., re-inspection, new tests), with 35% of initial conclusions later falsified. In contrast, objective-dominant cases (n = 59) required only 2.5 measures on average.

This finding underscores the necessity for the rigorous Context Information Modeling (Step 2) to neutralize bias and standardize heterogeneous inputs before causal reasoning.

Step 2: Context Information Modeling

Objective Fact Standardization

Field facts underpin fire causation reasoning; their completeness, objectivity, and structure directly determine the scientific rigor of downstream hypothesis generation. To convert unstructured inputs—such as scene notes, statements, and testimonies—into computable form, we separate fact standardization from causal judgment.

First, all input texts are classified into objective records (e.g., forensic reports, instrument logs) and subjective accounts (e.g., interviews, recollections), each processed under differentiated protocols:

- For objective records—which are reproducible and verifiable—the focus is on structural conversion with minimal alteration to original phrasing.
- For subjective accounts, no truth assessment is performed; instead, vague expressions are neutralized through standardized formatting.

Second, four specific textual quality issues are systematically addressed:

- Semantic vagueness (e.g., "possibly short-circuited," "about half an hour ago"),
- Anaphoric ambiguity (e.g., "that device," "the aforementioned location" without spatial anchor),
- Redundant entanglement (core facts intermixed with repetitive or irrelevant details),
- Subjective framing (e.g., "I think it started there").

Standardization rules include:

- Converting hedging terms ("probably") into attributed quotations (e.g., "According to Witness A…");
- Resolving ambiguous references via explicit contextual grounding (e.g., revising "there was burnt wiring" to "burnt wiring was observed at the northwest corner of the living room sofa");
- Removing redundant or emotionally charged content while preserving factual kernels;
- Merging consistent multi-source descriptions of the same event.

Critically, the system must not assess the credibility of subjective statements nor infer causal links between independent facts. Its function is strictly confined to neutral transcription and linguistic normalization.

Core Element Structured Extraction

Based on standardized objective fact descriptions, structured extraction proceeds according to the six core elements identified previously (timeline, spatial topology, equipment entities, combustibles, personnel, trace

characteristics), systematically mapping discrete fact units to analytical dimensions and establishing intra-dimensional and cross-dimensional associations. Each element's modeling aims to address a core question type in fire investigation; structured fields are detailed in Table 2.

Table 2. Fire Accident Structured Element Modeling: Elements and Specifications

Element Name	Core Definition	Extraction Fields
Timeline	Temporal coordinates and logical association chains of events	*Time point/period, behavior/event, involved personnel/equipment, occurrence location/position, other information*
Spatial Topology	Spatial layout and geometric relationships related to fire	*Area, position, functional use, structure, burn damage degree, special traces (soot, collapse, etc.), ventilation conditions, other information*
(Potential) Equipment Entities	Technical systems and energy carriers at fire scene	*Equipment model/category, user/manager, position, status, known risk points (maintenance history, inherent defects), other information*
(Potential) Combustibles	Material basis of fire load	*Category, item, location, status, known risk points, other information*
Personnel	Personnel activities and organizational factors before and after fire	*Personnel role, position, specific behavior/action, interaction with other personnel/equipment, status, other information*
Physical Evidence of Fire Pattern	Physical-chemical features and clues of fire residues	*Area/component, position, direction, trace type, morphological characteristics, fire direction indicators, other information*

Element extraction is not mere categorization but context-aware associative validation. For instance, the fact "P1 operated welding equipment in Area A at Time T_1" simultaneously links four dimensions—timeline (T_1), personnel (P1's action), equipment (welding device status), and spatial topology (Area A)—with contextual details enriched in each. When assigning facts to elements, consistency checks are enforced: if "welding equipment" is labeled "faulty," the system verifies whether objective evidence exists (e.g., maintenance logs or multiple witness reports of abnormal sounds); otherwise, it flags the claim as "to be verified." Cross-dimensional contradictions (e.g., witness statements vs. surveillance timestamps) are explicitly annotated as key questioning points for downstream reasoning.This process transforms heterogeneous texts into a structured, bias-mitigated factual foundation for subsequent hypothesis generation.

Step 3: Hypothesis Generation via Causal Matrix

Matrix Design

As discussed in background, both the Fire Triangle model and the MMEM framework exhibit explanatory blind spots when applied in isolation. To synergistically integrate these complementary perspectives, this study proposes a Fire Triangle–MMEM cross-matrix: a two-dimensional structure that systematically couples three physical dimensions—Fuel (F), Oxidizer (O), and Ignition Source (I)—with four organizational dimensions—Man (M), Machine (M), Environment (E), and Management (M)—yielding 12 causal cells.

Formally, let the Fire Triangle be represented as the set

$$\mathcal{F} = \{F_{\text{Fuel}}, F_{\text{Oxidizer}}, F_{\text{Ignition source}}\}$$

denoting the three fundamental physical prerequisites for fire. Let the MMEM framework be defined as

$$\mathcal{M} = \{M_{\text{Man}}, M_{\text{Machine}}, M_{\text{Environment}}, M_{\text{Management}}\}$$

representing systemic factors influencing accident evolution. The hypothesis space $\mathcal{H}$ is then constructed as the Cartesian product:

$$\mathcal{H} = \mathcal{F} \times \mathcal{M} = (f_i, m_j) \mid f_i \in \mathcal{F}, m_j \in \mathcal{M}$$

Each cell (f_i, m_j) represents a concrete causal mechanism. For example:

- ($F_{\text{Ignition Source}}$, M_{Man}): "Unauthorized human action generates unintended ignition energy" (e.g., illicit welding);
- (F_{Fuel}, $M_{\text{Environment}}$): "Environmental conditions expand the exposure range of combustibles" (e.g., wind-driven accumulation of flammable dust).

Table 3. Fire Causation Cross-Matrix: Fire Triangle × MMEM

	Man (M)	Machine (M)	Environment (E)	Management (M)
Fuel (F)	M–F *e.g., personnel introduce fuel*	M–F *e.g., equipment leakage*	E–F *e.g., environmental accumulation*	M–F *e.g., poor storage control*
Oxidizer (O)	M–O *e.g., unauthorized ventilation*	M–O *e.g., oxygen generator fault*	E–O *e.g., ambient O_2 concentration*	M–O *e.g., inadequate ventilation design*
Ignition Source (I)	M–I *e.g., smoking, hot work*	M–I *e.g., electrical short circuit*	E–I *e.g., lightning, spontaneous ignition*	M–I *e.g., failed hot - work permit system*

Each cell represents a concrete causation logic. For instance, "Personnel (M)-Ignition Source (I)" denotes "ignition source directly or indirectly caused by personnel factors." This systematic structure ensures exhaustive hypothesis generation while reducing omission risk and enhancing scientific rigor.

Hypothesis Instantiation

Instantiation of matrix unit (f_i, m_j) requires binding with specific elements from context information modeling results and raw investigation materials in the preceding section, anchoring at least one dimensional element instance to ensure factual support and verifiability of subsequent hypotheses. For example, for unit ($F_{Ignition\ Source}$,M_{Man}), the instantiation must simultaneously associate with specific operators in the "personnel" dimension (e.g., "Party A") and ignition carriers in the "equipment/trace" dimension (e.g., "electric heater" or "molten copper wire"), forming a complete and traceable causation pathway. Reference composition information for each unit is detailed in Table 4.

Table 4. Mapping Guidelines for Matrix Cell Instantiation

Matrix Unit $(\mathbf{f_i}, \mathbf{m_j})$	Typical Mechanism	Typical Causation Pattern	Priority Mapping Elements
(F_{fuel}, M_{man})	Personnel behavior causing abnormal fuel configuration	*e.g., non-compliant stacking, excess storage, incompatible material co-storage*	*Fuel + Personnel*
$(F_{fuel}, M_{machine})$	Equipment factors causing fuel state changes	*e.g., leakage, damage, temperature control failure*	*Fuel + Equipment Entities*
$(F_{fuel}, M_{environment})$	Environmental conditions affecting fuel combustibility	*e.g., temperature/humidity elevation, ventilation accumulation, oxidation acceleration*	*Fuel + Spatial Topology*
$(F_{fuel}, M_{management})$	Management defects causing fuel risk loss of control	*e.g., procurement approval absence, storage specification deficiency, inspection failure*	*Fuel + Physical Evidence of Fire Pattern*
$(F_{oxidizer}, M_{man})$	Personnel behavior altering oxidation conditions	*e.g., non-compliant ventilation opening, sealed environment destruction*	*Spatial Topology + Personnel*
$(F_{oxidizer}, M_{machine})$	Equipment factors causing oxygen enrichment/oxidizer release	*e.g., oxygen generation equipment malfunction, oxidizer container damage*	*Equipment Entities + Spatial Topology*
$(F_{oxidizer}, M_{environment})$	Environmental natural factors altering oxygen concentration	*e.g., pressure changes, ventilation system failure*	*Spatial Topology*
$(F_{oxidizer}, M_{management})$	Management defects causing oxidation condition loss of control	*e.g., ventilation design defects, change management deficiency*	*Spatial Topology + Physical Evidence of Fire Pattern*
$(F_{ignition\ source}, M_{man})$	Personnel behavior directly generating ignition energy	*e.g., non-compliant fire use, smoking, static discharge*	*Personnel + Equipment Entities + Timeline*

Matrix Unit (f_i, m_j)	Typical Mechanism	Typical Causation Pattern	Priority Mapping Elements
$(F_{ignition\ source}, M_{machine})$	Equipment malfunction releasing ignition energy	*e.g., electrical short circuit, mechanical friction overheating, static accumulation*	*Equipment Entities + Physical Evidence of Fire Pattern*
$(F_{ignition\ source}, M_{environment})$	Environmental natural factors providing ignition energy	*e.g., lightning, spontaneous ignition, high-temperature exposure*	*Spatial Topology + Physical Evidence of Fire Pattern*
$(F_{ignition\ source}, M_{management})$	Management defects causing ignition condition formation	*e.g., fire permit absence, equipment maintenance deficiency, training insufficiency*	*Personnel + Equipment Entities + Physical Evidence of Fire Pattern*

Using this schema, abstract causal mechanisms are grounded in structured evidence. The instantiation process performs a dual validation:

1. *Physical feasibility check: Do Fuel, Oxidizer, and Ignition coexist in space and time?*
2. *Organizational plausibility check: Is there supporting evidence of MMEM-level enabling conditions?*

This yields an initial hypothesis set $\mathcal{H}$ that jointly satisfies physical necessity and organizational sufficiency. For example:

H1: *Due to long-term neglect of electrical maintenance by property management (Management failure), an aged kitchen socket short-circuited under load (Equipment failure), igniting nearby polyurethane foam insulation (Fuel), while occupants were in deep sleep (Human factor), delaying detection and response.*

By design, this approach guarantees logical completeness and evidentiary anchoring in the initial hypothesis space, effectively mitigating investigation blind spots caused by cognitive biases or experiential limitations.

Step 4: Adversarial Validation via Multi-Agent Game

A core challenge in fire investigation lies in discriminating—among multiple logically coherent competing hypotheses—the explanation that best aligns with physical laws and empirical evidence. Traditional approaches relying on a single expert are vulnerable to cognitive biases, while group deliberations among investigators may succumb to authority dominance or groupthink, suppressing dissenting viewpoints. To address this, we introduce a non-cooperative, incomplete-information game-theoretic framework comprising specialized agents that engage in structured adversarial debate. Through an iterative "claim–challenge–response–adjudication" protocol, the system enables dynamic belief updating and robustness-based hypothesis selection.

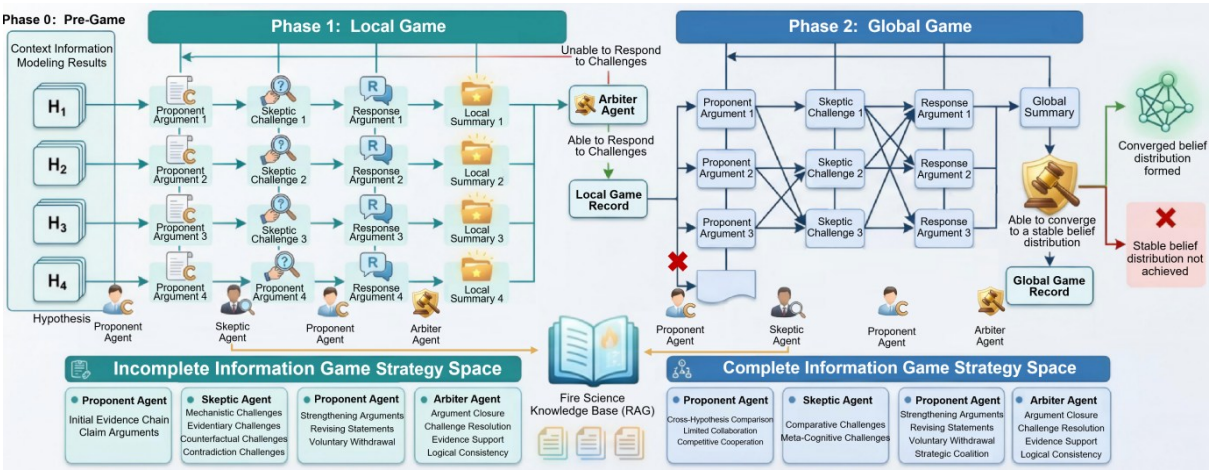

Figure 2. Two-phase dynamic game mechanism and process

Agent Roles and Strategies

The framework instantiates three categories of specialized agents using large language models, each operating within a non-cooperative game-theoretic structure and exercising autonomous decision-making.

- **Proponent Agent (A_i^P):** Exclusive agent for each hypothesis H_i, with core functionality of constructing evidence chains and responding to challenges. Strategy set includes: strengthening argumentation (deepening evidence chain details), revising formulation (adjusting secondary conditions to repair logical loopholes), strategic alliance (coordinating with other proponent agents to challenge common threats), and active abandonment (reducing defense intensity when fundamental evidence contradictions exist). All strategies employ context modeling results as the sole factual source, prohibiting introduction of unobserved information.
- **Skeptic Agent (A^S):** Not affiliated with any specific hypothesis, with core function of generating systematic challenges based on professional domain knowledge. Challenge types include: mechanism challenge (examining consistency with combustion science, electrical science, and other scientific principles), evidence challenge (scrutinizing evidence chain completeness and causal association strength), counterfactual challenge (proposing "what unobserved phenomena should occur if this hypothesis holds" to test predictive capability), and contradiction challenge (identifying logical conflicts between evidence pieces or within statements). Through RAG technology, dynamic access to external knowledge bases (fire science, forensic identification standards, etc.) ensures challenges are evidence-based, yet consistency verification with context modeling results is required to avoid pure theoretical deduction detached from field facts.
- **Arbiter Agent (A^A):** Serving as neutral adjudicator, maintaining game order and dynamically updating belief distributions. Assessment dimensions include: argumentation closure (whether Fire Triangle and MMEM elements are comprehensively covered), challenge resolution (whether key challenges are adequately addressed), evidence support (whether core assertions are directly and specifically supported), and logical consistency (whether fundamental conflicts exist with confirmed objective facts). Adjudication strategies: prioritize (significantly increase belief weight), requires further investigation (list information to be supplemented), or exclude (downgrade to secondary possibility).

Two-Phase Dynamic Game Mechanism

To balance reasoning depth with computational efficiency, the framework employs a two-phase progressive game (see Figure 2):

- **Phase 1: Local Game.** Each $\boldsymbol{A_i^P}$ argues solely based on its own hypothesis $\boldsymbol{H_i}$ and associated evidence subset, without awareness of competing hypotheses. $\boldsymbol{A^S}$ generates systematic challenges against the current hypothesis, with $\boldsymbol{A_i^P}$ responding individually. After each round, $\boldsymbol{A^A}$ outputs local vulnerability summaries based on four-dimensional assessment criteria, forming stage belief distribution $\boldsymbol{B}^{(l)}$. This phase aims to rapidly identify and eliminate candidate hypotheses with fundamental defects.
- **Phase 2: Global Game.** Information becomes fully public; all $\boldsymbol{A_i^P}$ agents access complete records and vulnerability summaries of all competing hypotheses, enabling optimization of arguments and strategic challenges against others. $\boldsymbol{A^S}$ upgrades challenge strategies, initiating comparative challenges(comparing explanatory power of different hypotheses for identical evidence) andmeta-cognitive challenges(questioning evidence set sufficiency or identifying anomalous phenomena). $\boldsymbol{A^A}$ introducesconsensus formation degreeandexplanatory parsimonyas auxiliary dimensions, dynamically updating belief distribution $\boldsymbol{B}^{(t)}$.

Game termination condition issemantic stability: belief distribution remains unchanged for two consecutive rounds, key challenges against the leading hypothesis are fully addressed, and no newdisruptive challenges emerge. At this point, $\boldsymbol{A^A}$ terminates the game, forming final belief distribution $\boldsymbol{B}^{(t)}$ and complete game archives.

Output and Downstream Application

Final outputs include:converged belief distribution $\boldsymbol{B}^{(t)}$ andcomplete game archives(recording argumentation evolution, key challenges, and response strategies). Downstream tasks encompass: outputting standardized causation determination opinions, proposing follow-up investigation recommendations for unresolved challenges, and organizing game archives into reusable knowledge assets.

EVALUATION

Experimental Validation

To evaluate MAGR-FI, we constructed a standardized test set of 1,051 cases derived from real-world

investigations and professional exams, covering diverse causation types (e.g., electrical faults, spontaneous ignition). Each case was scored on a 10-point scale based on standard rubrics. Using a controlled design, we compared the full MAGR-FI framework against baseline large language models using zero-shot prompting on raw case texts.

Table 5. Performance Comparison of MAGR-FI and Baseline Models in Fire Cause Investigation

Model Configuration	Average Score	Hard Case Performance (Hard 3%)	Avg. Game Rounds	Baseline Model Causation Hit Rate	Initial Hypothesis Set Recall	Optimal Hypothesis Hit Rate
Qwen3-32B (non-reasoning)	5.46	0.55	/	44.5%	/	/
Qwen3-32B (reasoning)	6.05	0.72	/	52.8%	/	/
Qwen3-32B (non-reasoning, continual pre-training)	5.92	0.89	/	48.1%	/	/
Qwen3-32B (reasoning, continual pre-training)	6.58	1.03	/	57.2%	/	/
Qwen3-32B (non-reasoning, MAGR-FI)	**6.72**	**1.35**	**4.21**	/	**87.4%**	**61.8%**
Qwen3-Max (non-reasoning)	6.95	1.78	/	59.7%	/	/
Qwen3-Max (reasoning)	7.46	2.25	/	67.4%	/	/
Qwen3-Max (non-reasoning, MAGR-FI)	**8.51**	**3.92**	**3.36**	/	**97.6%**	**80.2%**

Experimental results show that MAGR-FI–enhanced models (Qwen3-32B and Qwen3-Max) significantly outperform their baselines in both average score and hard-case performance. Qwen3-32B improved from 5.46 to 6.72 (average) and 0.55 to 1.35 (hard cases); Qwen3-Max saw even larger gains—8.51 vs. 6.95 (average) and 3.92 vs. 1.78 (hard cases)—demonstrating the framework's universality in boosting complex reasoning.

To dissect the improvement mechanism, we introduced three metrics:

(1) Baseline hit rate: proportion where the model's top hypothesis matches the true cause;

(2) Initial hypothesis recall: whether the true cause appears in the initial hypothesis set;

(3) Optimal hypothesis hit rate: whether the final top hypothesis after multi-agent debate matches the truth.

Recall reflects generation completeness; hit rate reflects validation discrimination—jointly evaluating framework integrity.

Results confirm strong gains in both stages:

- *Initial recall rose to 87.4% (Qwen3-32B) and 97.6% (Qwen3-Max);*
- *Final optimal accuracy reached 61.8% and 80.2%, respectively.*

This validates the two core designs: The Fire Triangle–MMEM causation matrix ensures critical pathways are not missed; Multi-agent adversarial gaming effectively discriminates among plausible hypotheses.Moreover, framework models converged in fewer rounds on average (4.21 and 3.36), indicating not only higher accuracy but also greater efficiency.

In summary, embedding LLMs within a domain-grounded, structured reasoning framework effectively mitigates their weaknesses in professional tasks—such as logical leaps, underutilized evidence, and instability—offering a robust pathway toward reliable, interpretable, and high-fidelity intelligent decision-making in fire investigation.

In-depth Case Analysis

To further reveal the framework's reasoning mechanisms in complex real-world scenarios, a comparative analysis was conducted on a real simulated rural residential fire case.

Core case information: *A rural residential fire resulted in deaths of the female homeowner and child, with homeowner Wang uninjured. Key scene inspection findings: lower portion of wooden bed at fire origin showed heavier burning than upper portion; gasoline flow combustion traces confirmed by testing on cement floor before bed; no deflagration damage in room; no electrical equipment except lighting fixture. Wang's testimony: due to marital argument and emotional agitation, he poured gasoline on floor before bed and ignited it with a*

non-extinguished cigarette butt.

The baseline model (Qwen3-Max with reasoning) focused on the superficial "gasoline + cigarette butt" evidence chain, concluding "intentional gasoline ignition" based on the cigarette's temperature (200–300°C) exceeding gasoline's flash point (−50°C), liquid-like burn patterns, and no explosion traces. However, it ignored critical physical constraints: in low-temperature, low-volatility conditions, gasoline vapor rarely reaches ignitable concentrations, and smoldering cigarette butts have low heat release rates, making ignition highly improbable.

MAGR-FI first generated three hypotheses:

- *H_1: cigarette-ignited vapor;*
- *H_2: lighting fixture spark;*
- *H_3: intentional arson using an open flame (cigarette claim false).*

For H_1, the Proponent cited temperature and burn morphology, while the Skeptic challenged:

(1) Physical feasibility: Freshly poured gasoline yields vapor either too rich (>7.6%) or too lean (<1.4%); ignition requires 30–120 seconds for concentration stabilization—unattainable with a smoldering butt.

(2) Ignition mechanism: "Tossing" a butt cannot sustain contact with the vapor layer near the floor; effective ignition would require deliberate "crouching placement."

(3) Behavioral inconsistency: Impulsive anger favors rapid action, yet successful ignition demands waiting—contradicting the stated emotional state.

In the global game, the Skeptic even questioned the confession's authenticity. The Arbiter concluded: H_1 lacks physical plausibility; H_3 is motivationally coherent but evidentially weak; thus, Wang's account is partially truthful, but the "cigarette ignition" claim is physically implausible and likely a cover-up for deliberate arson.

This case shows MAGR-FI not only evaluates surface-plausible hypotheses but exposes their hidden fragilities through adversarial debate—shifting reasoning from single explanation to competitive, self-correcting hypothesis evolution, significantly enhancing rigor, robustness, and interpretability.

CONCLUSIONS AND DISCUSSION

Summary

This study proposes MAGR-FI, a structured reasoning framework that integrates fire science theory with multi-agent game mechanisms to improve accuracy and interpretability in fire cause attribution. By constructing a Fire Triangle–MMEM cross-matrix for systematic hypothesis generation and deploying a Proponent–Skeptic–Arbiter tri-agent architecture for adversarial validation, the framework formalizes the implicit deliberation of human experts. On a test set of 1,051 real-world cases, it significantly outperforms direct LLM prompting—especially in complex, ambiguous scenarios—demonstrating greater robustness and reasoning depth. Case analysis confirms its ability to detect physically implausible yet superficially coherent explanations, expose testimonial inconsistencies, and highlight evidentiary weaknesses, leading to more reliable conclusions.By transforming retrospective causal analysis into structured, auditable knowledge assets, MAGR-FI closes the feedback loop between incident investigation and organizational learning, thereby supporting proactive risk prevention and enhanced crisis governance.

Rethinking AI for Crisis Reasoning

MAGR-FI transcends the traditional view of LLMs as mere tools for information extraction or pattern matching, transforming them from passive statistical learners into active theory-consistent reasoners. Against growing concerns over LLM hallucinations, hidden biases, and accountability in high-stakes decisions, MAGR-FI generates auditable multi-agent game-theoretic archives, effectively addressing the explainability deficit and attribution challenges in critical AI applications. For serious domains such as law, these transparent decision logs provide the necessary evidentiary basis for accountability, ensuring AI outputs are no longer untraceable black-box results but verifiable, defensible logical derivations.

Notably, MAGR-FI abandons the pursuit of full AI automation and instead advocates a process-centric human-AI collaboration paradigm. **In practical applications (e.g., fire investigation), the value of AI lies not in providing a "final answer," but in systematically revealing its reasoning trajectory—such as explicitly presenting rejected hypotheses, evidentiary contradictions, and gaps in information.** This design enables

three forms of genuine collaboration: (1) targeted supplementary investigation, where system-flagged "to-be-verified" nodes become prioritized on-site checklists; (2) real-time intervention, allowing investigators to correct extraction biases or eliminate implausible hypotheses at any stage; (3) organizational learning, whereby novice investigators internalize expert-style skepticism and causal reasoning through adversarial debate archives, facilitating tacit knowledge externalization.

Although instantiated in fire investigation, MAGR-FI's design rationale has cross-domain transfer potential. Analytical tasks requiring theory-guided causal attribution under incomplete information—such as industrial accident investigation or public health event tracing—can draw upon the core principles of this framework.

Limitations and Future Work

Limitations remain: (1) performance depends on input report quality; omissions or inaccuracies constrain outputs; (2) multi-agent computation is costly, requiring efficiency optimization for scale; (3) agent interaction depth is bounded by underlying LLM capabilities; (4) the framework currently addresses only causation reasoning, not upstream tasks like scene inspection or forensic commissioning.

Future work will: (1) integrate multimodal data (e.g., images, sensor logs) to strengthen factual grounding; (2) extend to chemical and mining disasters to test generalization; (3) build interactive interfaces enabling real-time investigator intervention, realizing true human-machine co-intelligence and advancing fire investigation into a new era of theory-guided, data-augmented, collaborative analysis.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of Beijing (Grant No. L255011, 8242014), the National Natural Science Foundation of China (Grant No. 72521001), the Chinese Academy of Engineering Local Cooperation Project (Grant No. 2025-AHYJY-06), and Strategic Study Project of Chinese Academy of Engineering (Grant No. 2023-JB-08). The authors sincerely acknowledge their support.

REFERENCES

Liu, J., Yang, G., Wang, W., Zhou, H., Hu, X., & Ma, Q. (2022). Based on ISM—NK Tunnel Fire Multi-Factor Coupling Evolution Game Research. *Sustainability*, 14(12), 7034.

Zhu, D., Ding, J., & Han, X. (2025, August). The mechanism of multi-factor coupling of fire and explosion accidents of hazardous chemicals based on complex network model. *In Journal of Physics: Conference Series (Vol. 3092, No. 1, p. 012025)*. IOP Publishing.

Okoli, J. O., Weller, G., & Watt, J. (2016). Information processing and intuitive decision-making on the fireground: towards a model of expert intuition. *Cognition, Technology & Work*, 18(1), 89-103.

Iliadis, L. S., Papastavrou, A. K., & Lefakis, P. D. (2002). A heuristic expert system for forest fire guidance in Greece. *Journal of environmental management*, 65(3), 327-336.

Sankarasubramanian, P., & Ganesh, E. N. (2020). Fire investigation and assessment using CNN and image processing. *Journal of Critical Reviews*, 7(19), 9825-9830.

Ingle, P. Y., & Gab-Kim, Y. (2025, October). FireNarrator: Multimodal LLM-Based Fire Incident Reporting with Decision Logic. In *2025 IEEE Conference on Dependable, Autonomic and Secure Computing (DASC)* (pp. 65-71). IEEE.

Hine, G. A. (2004). Fire scene investigation: an introduction. *Analysis and interpretation of fire scene evidence*, 33.

Lentini, J. J. (2018). *Scientific protocols for fire investigation*. CRC press.

Munday, J., & Gardiner, M. (2013). Fire investigation policies and practices in the UK. In *Handbook of Forensic Science* (pp. 254-277). Willan.

Xu, K. (2024, July). Pathway and Empirical Study of Fire Accident Intelligence Support Based on Case Reasoning and Event Logic Graph. In *2024 20th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)* (pp. 1-6). IEEE.

Goh, Y. M., & Chua, D. K. H. (2010). Case-based reasoning approach to construction safety hazard identification: Adaptation and utilization. *Journal of Construction Engineering and Management*, 136(2), 170-178.

Liu, J. (2009, November). Case-based reasoning intelligent decision approach for firefighting tactics. In *2009 Second International Conference on Intelligent Networks and Intelligent Systems* (pp. 437-440). IEEE.

Chandra, R., Agarwal, S., & Tiwari, S. (2025, August). Ontology-Based Forest Fire Management Using Complex Event Processing and Large Language Models. In *International Conference on Database and Expert Systems Applications* (pp. 98-112). Cham: Springer Nature Switzerland.

Zhao, S., Prapas, I., Karasante, I., Xiong, Z., Papoutsis, I., Camps-Valls, G., & Zhu, X. X. (2024). *Causal graph neural networks for wildfire danger prediction*. arXiv preprint arXiv:2403.08414.

Mirończuk, M. M. (2020). Information extraction system for transforming unstructured text data in fire reports into structured forms: a Polish case study. *Fire technology*, 56(2), 545-581.

Yan, H., Ma, X., Chen, F., Zhao, R., & Jia, L. (2021, October). Knowledge modeling and analysis for railway fire accident using ontology-based knowledge graph. In *International Conference on Electrical and Information Technologies for Rail Transportation* (pp. 573-591). Singapore: Springer Singapore.

Zhang, L., & Ashley, K. D. (2025). *Mitigating manipulation and enhancing persuasion: A reflective multi-agent approach for legal argument generation*. arXiv preprint arXiv:2506.02992.

Nimbalkar, M., Chandre, P., Shendkar, B., Jagdale, S., Arbat, R., & Dhopte, S. (2025, December). Agent AI for Personalized Healthcare: A Multi-Agent Framework for Real-Time Disease Detection and Patient Support. In *2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG)* (pp. 1-8). IEEE.

Chahine, C., Vidal, T., El Falou, M., & Pérès, F. (2022). Multi-Agent Dynamic Planning Architectures for Crisis Rescue Plans. In *ISCRAM* (pp. 243-255).

Drysdale, D. (2011). *An introduction to fire dynamics*. John wiley & sons.

Reason, J. (1990). *Human error*. Cambridge university press.

Du, Z., Qian, C., Liu, W., Xie, Z., Wang, Y., Qiu, R., ... & Han, L. (2025). *Multi-agent collaboration via cross-team orchestration*. arXiv.

Qian, C., Xie, Z., Wang, Y., Liu, W., Zhu, K., Xia, H., ... & Sun, M. (2024). *Scaling large language model-based multi-agent collaboration. arXiv preprint* arXiv:2406.07155.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., ... & Schmidhuber, J. (2023, August). MetaGPT: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*.