Post-Hoc Knowledge Grounding for Verifiable Multi-Agent Search and Rescue Decision Support

Yayun Tan[1*], **Franz Kurfess**[1*], **Emmanuel Delgado**[1],
Christopher Young[2], **Gary Bloom**[3]

[1]California Polytechnic State University, San Luis Obispo
[2]Contra Costa County Sheriff's Search & Rescue
[3]San Mateo County Sheriff's Search & Rescue

Abstract

Multi-agent systems powered by large language models (LLMs) show promise for complex decision-support tasks, yet their outputs often contain hallucinations that undermine trust in safety-critical domains such as search and rescue (SAR). We present a post-hoc verification framework that grounds LLM agent outputs in a probabilistic knowledge graph derived from approximately 12,000 curated cases from the International Search and Rescue Incident Database (ISRID). Our approach extracts structured claims from agent outputs, performs probabilistic reasoning to detect statistical anomalies, and produces tiered decisions (accept, flag, reject) with explicit evidence chains. The framework operates downstream of a seven-agent SAR system, verifying terrain predictions (where the subject is likely found) and status predictions (the subject's likely condition). These claims directly inform search prioritization. Experiments demonstrate that coordinators following our system's recommendations achieve 71.0% accuracy, compared to 51.6% when accepting all LLM outputs, a 19.4 percentage point improvement. The system detects anomalies with 81.3% F1 score while maintaining practical coverage (46.3% of claims verified). Ablation studies confirm anomaly detection as the critical component, contributing +5.7 percentage points over grounding alone. Our framework provides a practical path toward trustworthy AI in emergency response.

1 Introduction

Search and rescue (SAR) operations are time-critical, high-stakes endeavors where incorrect decisions can cost lives. When a person goes missing, SAR coordinators must rapidly synthesize heterogeneous information (weather conditions, terrain analysis, witness interviews, historical patterns) to prioritize search areas and allocate resources (Young 2022). The complexity and urgency of these decisions make SAR an appealing domain for AI-assisted decision support.

Recent advances in large language models (LLMs) have enabled sophisticated multi-agent systems where specialized agents handle perception, reasoning, and synthesis tasks (Wu et al. 2023). However, deploying such systems in safety-critical domains faces a fundamental challenge: LLM agents frequently generate plausible-sounding but factually incorrect outputs, a phenomenon known as hallucination (Ji et al. 2023). In multi-agent settings, these errors can propagate and compound, leading to confidently wrong recommendations that are difficult to detect.

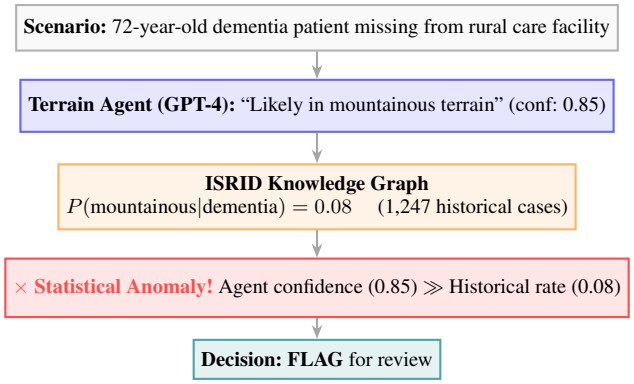

Figure 1. Post-hoc verification of a multi-agent SAR prediction. Our framework detects a statistical anomaly between agent confidence and historical probability.

We address this challenge through **post-hoc knowledge grounding**: verifying agent outputs against a structured knowledge base after generation, rather than attempting to prevent errors during generation. As illustrated in Figure 1, an LLM terrain agent may predict "mountainous terrain" with high confidence (0.85) for a dementia patient, but historical ISRID data from 1,247 cases shows this terrain occurs only 8% of the time. Our key insight is that such historical SAR data encodes statistical patterns that can serve as sanity checks for LLM predictions, enabling detection of hallucinations that would otherwise go unnoticed.

Our contributions are:

1. A post-hoc verification framework that grounds multi-agent LLM outputs in probabilistic knowledge graphs, enabling contradiction detection without modifying agent architectures.
2. A contradiction-aware decision mechanism that identifies claims conflicting with historical patterns and provides tiered recommendations with explicit evidence chains.
3. An empirical evaluation demonstrating 71.0% accepted

*Corresponding authors.

precision versus 51.6% for unverified LLM outputs (+19.4 percentage points) on realistic SAR scenarios derived from the International Search and Rescue Incident Database (ISRID) (Koester 2008).

Importantly, our *post-hoc* verification operates after agents generate predictions but before decisions are acted upon, enabling real-time correction during active SAR missions rather than retrospective analysis after mission completion.

2 Related Work

Multi-Agent LLM Systems. Frameworks like AutoGen (Wu et al. 2023), CAMEL (Li et al. 2023), and MetaGPT (Hong et al. 2023) enable LLM agents to collaborate on complex tasks. However, these systems inherit the hallucination tendencies of their underlying models, and errors can propagate through agent interactions. Prior work has explored self-consistency (Wang et al. 2022) and debate-based verification (Du et al. 2023), but these approaches rely on the same potentially unreliable models for verification. In contrast, our approach uses an external knowledge source (ISRID) that is independent of the LLM, providing verification based on approximately 12,000 curated historical cases rather than model-generated alternatives. This enables detection of systematic biases that self-consistency cannot catch, such as LLMs consistently overestimating survivability or underestimating terrain difficulty. Recent benchmarks (Liu et al. 2023) have highlighted the gap between LLM capabilities and reliable task execution, motivating external verification mechanisms.

Knowledge Grounding and Fact Verification. Retrieval-augmented generation (RAG) (Lewis et al. 2020) grounds LLM outputs by conditioning on retrieved documents. GraphRAG (Edge et al. 2024) extends this by constructing community-level summaries over knowledge graphs to improve retrieval for global queries, while StructGPT (Jiang et al. 2023) enables LLMs to reason over structured data including knowledge graphs. However, both RAG and GraphRAG operate during generation, requiring architectural modifications. Post-generation verification approaches like RARR (Gao et al. 2023) and Chain-of-Verification (Dhuliawala et al. 2023) edit outputs after generation but still rely on LLM-based fact-checking. FActScore (Min et al. 2023) decomposes generations into atomic facts for verification, while SAFE (Wei et al. 2024) uses search-augmented factuality evaluation. These methods excel at verifying factual claims with definitive answers ("Paris is the capital of France"), but struggle with probabilistic predictions ("A hiker is likely in forested terrain"). In contrast to RAG/GraphRAG approaches that inject knowledge *during* generation, our post-hoc approach verifies outputs *after* generation using probabilistic reasoning over domain-specific conditional distributions, enabling uncertainty quantification where no single "correct" answer exists.

Neurosymbolic AI and Verification. Hybrid approaches combining neural networks with symbolic reasoning have shown promise for reliable AI (Garcez and Lamb 2023). Knowledge graphs provide structured representations amenable to logical inference (Hogan et al. 2021). Probabilistic extensions enable uncertainty quantification (Pan et al. 2024). We extend this paradigm to multi-agent verification, using conditional probability tables derived from domain data to detect statistically implausible claims. Unlike logic-based verification that requires formal specifications ("a person cannot be in two places simultaneously"), our statistical approach operates on learned distributions from historical data, enabling soft constraints ("dementia patients are rarely found in mountainous terrain") that better match the uncertainty inherent in SAR predictions.

SAR Decision Support Systems. Traditional SAR planning relies on statistical models like the ISRID database (Koester 2008, 2023) and Mattson consensus tables for lost person behavior prediction. Early computational approaches include the MIST toolkit for missing person intelligence synthesis (Shaabani et al. 2016). The ISRID dataset has enabled data-driven approaches to estimate find locations based on subject category, terrain, and scenario, including agent-based models of lost person behavior (Hashimoto et al. 2022; Nguyen et al. 2023). Belden (2024) applied Bayesian networks to ISRID for causal inference in SAR outcomes, while Kim (2025) developed explainable AI methods to identify key predictors of search success. Young, Kurfess, and Bloom (2024) formalized the search intelligence process using AI, and Young, Kurfess, and Bloom (2025) demonstrated advanced data analysis techniques for missing person searches. Topographic image processing for path-finding in SAR contexts has been explored by Washington (2018). Recent work has also applied multi-robot coordination to SAR planning (Queralta et al. 2020). However, integration of LLM-based multi-agent systems with structured historical knowledge for holistic decision support remains underexplored. Our work bridges this gap by using historical SAR statistics as a verification layer for LLM-generated recommendations, building on the companion agentic SAR system described by Kurfess, Tan, and Delgado (2025).

3 Multi-Agent SAR System

Before describing our verification framework, we provide context on the multi-agent SAR system whose outputs we aim to verify (see Figure 2 for an overview). The system comprises seven specialized agents coordinated via Redis Streams using a publish-subscribe architecture.

3.1 Agent Architecture

The system comprises seven specialized agents organized into four functional categories (Kurfess, Tan, and Delgado 2025). The agents are heterogeneous: some rely on LLMs (GPT-4, Gemini), others on computer vision models or direct API access, and some are purely rule-based. All agents produce structured outputs with confidence scores, published to Redis Streams for downstream consumption.

Perception Agents: The *Weather Agent* retrieves real-time NWS forecasts (precipitation, temperature, visibility) via the NOAA API. The *Photo Analysis Agent* performs ob-

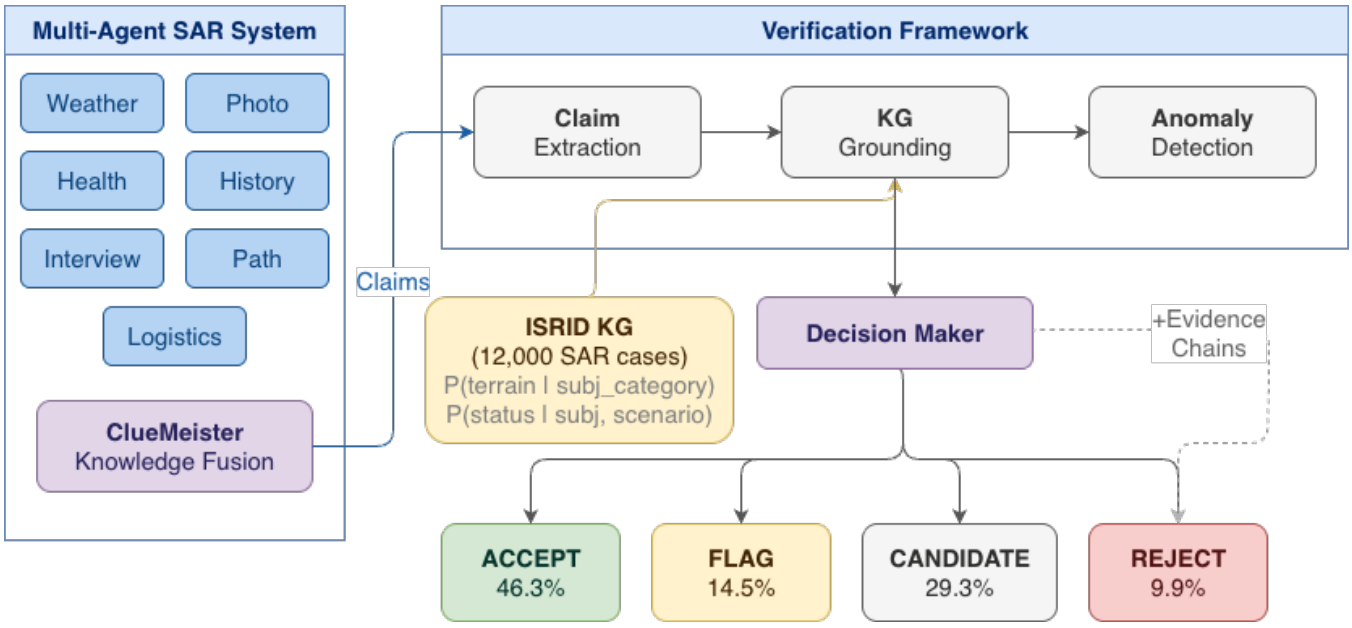

Figure 2. Post-hoc verification framework for multi-agent SAR systems. Seven specialized agents (left) produce predictions verified (right) through claim extraction, KG grounding, and anomaly detection.

ject detection on aerial and ground imagery for subject identification.

Knowledge Agents: The *History Agent* implements RAG over ISRID using vector search with sentence-transformer embeddings, retrieving similar historical cases to inform predictions. The *Interview Agent* extracts behavioral indicators from witness transcripts, and the *Health Agent* assesses medical risk factors for survivability estimation by synthesizing mission data, weather conditions, and field observations.

Planning Agents: The *Path Analysis Agent* analyzes DEMs and OpenStreetMap data for terrain-aware route identification (Washington 2018), while the *Logistics Agent* coordinates resource allocation from equipment and personnel inventories.

Orchestration: The *ClueMeister Agent* maintains a Neo4j knowledge graph aggregating all upstream outputs, correlating entities across sources and generating cross-agent insights. This agent serves as the critical integration point for our verification framework. The *Command Agent* coordinates agent interactions.

The current seven-agent configuration represents the initial deployment; the modular architecture supports extension with additional specialized agents as operational requirements evolve.

3.2 Verification Integration Point

Our post-hoc verification framework operates at the output of the ClueMeister agent, which aggregates recommendations from all upstream agents. Rather than modifying individual agent architectures, we intercept the fused recommendations and verify claims against historical patterns before presenting them to SAR coordinators. This design enables:

1. **Non-invasive integration:** Existing agents remain unchanged.
2. **Centralized verification:** All claims pass through a single verification checkpoint.
3. **Graceful degradation:** If verification is unavailable, the system falls back to unverified recommendations.

3.3 Scope of Verification

While the multi-agent system produces diverse outputs (weather forecasts, photo annotations, interview summaries, logistics plans), our verification framework focuses on claims that can be statistically grounded in historical SAR data. Specifically, we verify:

- **Terrain predictions:** Where the subject is likely to be found (e.g., mountainous, flat, water).
- **Status predictions:** The subject's likely condition (e.g., well, injured, deceased).

This scope is determined by the structure of the ISRID database, which contains rich historical statistics on terrain and status outcomes across 37 subject categories. Other agent outputs (weather data, photo detections, resource allocations) serve as *contextual inputs* that inform terrain and status predictions, but are not themselves subject to probabilistic verification against ISRID. For example, the Weather Agent's forecast may influence a status prediction ("cold temperatures increase hypothermia risk"), but we verify the resulting status claim rather than the weather data itself.

This scoping decision reflects both practical and methodological constraints: ISRID provides comprehensive historical statistics for terrain and status outcomes, while other

agent outputs (real-time weather, photo detections) lack comparable historical benchmarks for probabilistic grounding. We prioritize verifying predictions that (1) directly inform search prioritization decisions and (2) can be statistically grounded in historical patterns with sufficient sample sizes.

4 Verification Framework

4.1 Problem Formulation

Given the multi-agent system described above, we formalize the verification problem. Let $\mathcal{A} = \{a_1, \ldots, a_n\}$ be a set of agents, and let O_i denote the output of agent a_i for a given scenario. Our goal is to verify the claims in $\bigcup_i O_i$ against a knowledge graph $\mathcal{G}$ derived from historical SAR data, producing:

1. A set of grounded claims with associated confidence scores
2. Detected contradictions between claims and historical patterns
3. Tiered decisions (accept, flag, candidate, reject) with evidence chains

Figure 2 illustrates our verification pipeline. Agent outputs are first processed to extract structured claims. Each claim is then grounded against the probabilistic knowledge graph using single-hop or two-hop reasoning. An anomaly detector identifies claims that conflict with historical patterns. Finally, a decision maker integrates grounding confidence and anomaly signals to produce tiered recommendations.

4.2 Probabilistic Knowledge Graph

We construct a probabilistic knowledge graph from the International Search and Rescue Incident Database (ISRID) (Koester 2008). While ISRID contains a larger corpus of incidents, many lack complete attribute data; we use approximately 12,000 curated cases with structured attributes including subject category (e.g., hiker, child, dementia patient), terrain type, scenario conditions, and outcome status.

The knowledge graph encodes conditional probability distributions:

- **Single-hop:** $P(\text{Terrain} \mid \text{Subject})$ and $P(\text{Status} \mid \text{Subject})$ capture direct relationships between subject types and likely outcomes.
- **Two-hop:** $P(\text{Status} \mid \text{Subject}, \text{Scenario})$ captures more nuanced relationships conditioned on both subject type and environmental factors.

In practical terms, these probabilities answer questions like: "Based on 1,247 historical cases involving dementia patients, what percentage were found in flat terrain?" (67%) or "Among hikers lost in mountainous areas, what fraction were found alive?" (89%). This enables verification grounded in real-world SAR experience rather than LLM-generated estimates.

We filter probability tables to include only combinations with at least k historical observations (we use $k = 30$, a standard threshold for reliable frequency estimation where the normal approximation becomes reasonable) to ensure statistical reliability. This yields 37 subject categories with single-hop distributions and 39 subject-scenario combinations with two-hop distributions.

For each query, probabilities are computed from frequency counts:

$$P(Y = y \mid X = x) = \frac{\text{count}(X = x, Y = y)}{\text{count}(X = x)} \quad (1)$$

Data Quality Considerations. ISRID spans several decades of SAR incidents, but data quality varies: not all records have complete attributes, some cases predate modern technologies (potentially shifting behavioral patterns), and voluntary reporting may introduce regional biases. Our minimum sample threshold ($k = 30$) ensures statistical reliability, while CANDIDATE status for ungrounded claims (29.3%) routes uncertain cases for human review.

4.3 Claim Extraction and Grounding

We extract structured claims from agent outputs using an LLM-based extractor that identifies factual assertions about:

- **Terrain predictions:** "The subject is likely in forested terrain"
- **Status predictions:** "The subject is probably alive and mobile"

Each claim records six fields: the source agent, the claim type (terrain or status), the subject type, the scenario type, the predicted value, and the agent's stated confidence in $[0, 1]$.

Grounding Process. Grounding proceeds as follows:

1. **Entity alignment:** Match the claim's subject to a category in the KG using fuzzy string matching with a similarity threshold of 0.8.
2. **Single-hop lookup:** Query $P(\text{value} \mid \text{subject})$ from the corresponding probability table.
3. **Two-hop fallback:** If scenario context is available and the subject-scenario combination exists in the KG, query $P(\text{value} \mid \text{subject}, \text{scenario})$ for more specific grounding.

A claim is considered *grounded* if a matching entry exists in the KG. The grounding confidence is the probability of the claimed value given the subject (and optionally scenario).

4.4 Statistical Anomaly Detection

We detect statistical anomalies (claims that deviate significantly from historical patterns) through three mechanisms:

Low Probability Detection. A claim is flagged as anomalous if its grounding probability falls below a threshold $\tau_{\text{low}} = 0.15$ (i.e., the predicted outcome occurs in fewer than 15% of historical cases for that subject category, rare enough to warrant scrutiny):

$$\text{LowProb}(c) = \mathbb{I}[P(\text{value} \mid \text{subject}) < \tau_{\text{low}}] \quad (2)$$

For example, if a hiker is predicted to be in "urban" terrain but $P(\text{urban} \mid \text{hiker}) = 0.03$, this is flagged as contradicting historical patterns.

Overconfidence Detection. When an agent's stated confidence significantly exceeds the KG probability:

$$\text{Overconf}(c) = \mathbb{I}[\text{conf}(c) - P(\text{value} \mid \text{subject}) > \tau_{\text{mismatch}}] \tag{3}$$

where $\tau_{\text{mismatch}} = 0.40$, meaning the agent's confidence exceeds the historical rate by at least 40 percentage points, indicating substantial overconfidence relative to the evidence.

Underconfidence Detection. When an agent underestimates a well-supported prediction (less critical but still flagged).

Each detected anomaly is assigned a severity score based on the magnitude of the probability deviation:

$$\text{severity}(c) = \begin{cases} 1 - \frac{P(\text{value}|\text{subject})}{\tau_{\text{low}}} & \text{if LowProb}(c) \\ \frac{\text{conf}(c) - P(\text{value}|\text{subject})}{\tau_{\text{mismatch}}} & \text{if Overconf}(c) \end{cases} \tag{4}$$

This severity score enables nuanced decision-making rather than binary accept/reject.

4.5 Verification Algorithm

Algorithm 1 presents the complete verification procedure. For each claim, we first attempt grounding via single-hop or two-hop lookup, then apply anomaly detection, and finally produce a tiered decision with an evidence chain.

4.6 Tiered Decision Mechanism

We integrate grounding confidence and contradiction signals into a tiered decision framework:

- **ACCEPT** (≥ 0.70 confidence, no strong contradiction): The claim is well-supported by historical data.
- **FLAG** (0.50–0.70 confidence or minor contradiction): Requires human review.
- **CANDIDATE** (0.30–0.50 confidence): Plausible but uncertain.
- **REJECT** (severity ≥ 0.70 or very low probability): Conflicts with historical patterns.

For each decision, we generate an *evidence chain* explaining the reasoning:

> "Claim: Hiker likely in forest. **ACCEPT** (confidence: 0.82). Historical support: P(forest | hiker) = 0.67 based on 2,341 cases. No contradictions detected."

This transparency enables SAR coordinators to understand and appropriately weight AI recommendations.

Algorithm 1 Post-Hoc Claim Verification

Require: Claim $c = (\text{type}, \text{subject}, \text{value}, \text{conf})$
Require: Knowledge Graph $\mathcal{G}$ with probability tables
Require: Thresholds $\tau_{\text{low}}, \tau_{\text{mismatch}}, \tau_{\text{accept}}, \tau_{\text{flag}}$
Ensure: Decision $d \in \{\text{ACCEPT, FLAG, CANDIDATE, REJECT}\}$
Ensure: Evidence chain E

$E \leftarrow \emptyset$ {Initialize evidence}
// Step 1: Grounding
$p_{\text{kg}} \leftarrow \text{TwoHop}(\mathcal{G}, c.\text{subj}, c.\text{scen}, c.\text{val})$
if $p_{\text{kg}} = \text{null}$ **then**
  $p_{\text{kg}} \leftarrow \text{SingleHopLookup}(\mathcal{G}, c.\text{subject}, c.\text{value})$
end if
if $p_{\text{kg}} = \text{null}$ **then**
  E.add("Claim not grounded in KG")
  return $(\text{CANDIDATE}, E)$
end if
E.add("Grounded: $P(\text{value}|\text{subject}) = p_{\text{kg}}$")
// Step 2: Anomaly Detection
anomaly $\leftarrow$ false; severity $\leftarrow 0$
if $p_{\text{kg}} < \tau_{\text{low}}$ **then**
  anomaly $\leftarrow$ true
  severity $\leftarrow 1 - p_{\text{kg}}/\tau_{\text{low}}$
  E.add("LOW_PROB: Hist. prob. low")
else if $c.\text{conf} - p_{\text{kg}} > \tau_{\text{mismatch}}$ **then**
  anomaly $\leftarrow$ true
  severity $\leftarrow (c.\text{conf} - p_{\text{kg}})/\tau_{\text{mismatch}}$
  E.add("OVERCONF: Conf exceeds KG")
end if
// Step 3: Tiered Decision
if severity ≥ 0.7 **then**
  return (REJECT, E)
else if $p_{\text{kg}} \geq \tau_{\text{accept}}$ **and not** anomaly **then**
  return (ACCEPT, E)
else if $p_{\text{kg}} \geq \tau_{\text{flag}}$ **then**
  return (FLAG, E)
else
  return $(\text{CANDIDATE}, E)$
end if

5 Experiments

5.1 Experimental Setup

Dataset. We use the ISRID dataset containing approximately 12,000 historical SAR incidents. We split the data 80/20 for KG construction (training) and scenario generation (testing). Test scenarios are sampled to create realistic SAR situations with known ground truth outcomes.

Agents. We implement two LLM-based agents using GPT-4: a *terrain agent* that predicts likely terrain types based on subject category and conditions, and a *status agent* that predicts likely outcomes (alive/deceased, mobile/immobile). Each scenario generates 2 claims (one per agent type), evaluated against ground truth from the test set.

Baselines. We compare against:

- **Raw LLM:** Accept all agent predictions without verification
- **Rule-based (0.5/0.7):** Accept claims with agent confidence above threshold
- **Self-Consistency** (Wang et al. 2022): Sample LLM multiple times and accept if majority agrees (simulated via confidence-based voting)
- **Retrieval:** Accept claims supported by $\geq$50% of similar historical cases (using KG probability as proxy)
- **Grounding Only:** KG grounding without anomaly detection

Metrics. We report metrics that directly measure practical value for SAR coordinators:

- *Accepted Precision:* If a coordinator only acts on claims our system accepts, what fraction are actually correct? This is the primary metric: $\text{Prec}_{\text{accept}} = \frac{\text{correct accepts}}{\text{total accepts}}$.
- *Coverage:* What fraction of all claims does the system accept? This measures practical utility; too low coverage means too many cases require manual review.
- *Reject Precision:* When the system rejects a claim, how often was the claim actually wrong?
- *Anomaly F1:* Precision/recall for detecting predictions that deviate from historical probability distributions.

Protocol. We evaluate on 100 test scenarios (200 claims) with 5 random seeds, reporting mean ± standard deviation.

5.2 Main Results

Table 1 presents our main results. *Accepted Precision* measures: if coordinators act only on accepted claims, what percentage are correct? *Coverage* is the proportion of claims receiving ACCEPT. The key question is: *if SAR coordinators only act on claims our system accepts, how much does their decision accuracy improve?*

The full pipeline achieves **71.0% accepted precision**, meaning 71% of accepted claims are actually correct. This represents a **+19.4 percentage point improvement** over the 51.6% baseline of accepting all raw LLM outputs, while maintaining reasonable coverage (46.3% of claims accepted). Figure 3 visualizes the precision-coverage tradeoff across methods.

Why confidence thresholding fails. Raw LLM, Rule-based (0.5), and Rule-based (0.7) achieve identical precision (51.6%) because they accept virtually all claims. This is not a bug; it reflects a well-known property of LLMs: *systematic overconfidence*. Our LLM agents report confidence scores predominantly in the 0.75–0.95 range regardless of actual prediction accuracy. Consequently, thresholds below 0.75 reject almost nothing, demonstrating that simple confidence filtering is ineffective for overconfident models. Only at threshold 0.90 does filtering begin to work, but at severe cost to coverage (25.5%).

Self-consistency (51.1%) offers no improvement because LLMs reproduce the same errors consistently across samples: if an LLM overestimates survivability, it does so reliably. Retrieval-based verification (63.0%) and Grounding Only (65.3%) perform better by leveraging historical patterns, but without explicit anomaly detection they cannot identify predictions that are statistically implausible. Adding anomaly detection yields an additional **+5.7%** improvement over grounding alone (65.3% → 71.0%), demonstrating that identifying statistically implausible claims is crucial for reliable verification.

Understanding Coverage. Our full pipeline ACCEPTs 46.3% of claims; the remaining 53.7% are distributed as: 14.5% FLAG (requires human review), 9.9% REJECT, and 29.3% CANDIDATE (insufficient KG coverage). Crucially, *all 100% of claims are triaged*. This conservative behavior is *by design*: in safety-critical domains, the cost of acting on a wrong prediction far exceeds the cost of routing a claim for human review. Figure 3 shows the precision-coverage tradeoff: confidence thresholding achieves comparable precision only at much lower coverage (Rule-based 0.90: 53.9% at 25.5% coverage vs. our 71.0% at 46.3% coverage), confirming our approach's superior operating point.

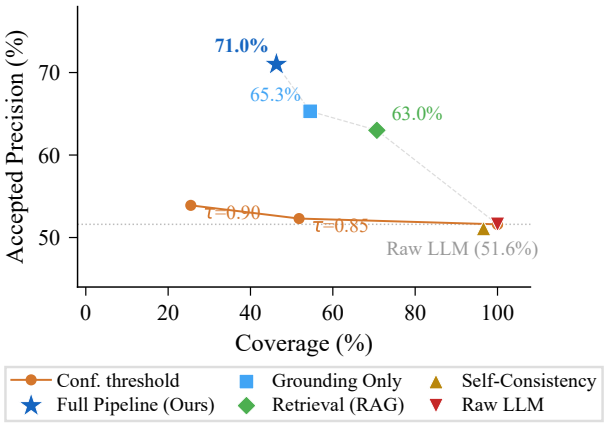

Figure 3. Precision-coverage tradeoff across methods. Our full pipeline achieves the best balance of precision and coverage.

5.3 Detailed Performance Analysis

Table 2 provides additional performance details. The system's 71.0% accepted precision means that when coordinators follow our system's ACCEPT recommendations, they will be correct 71% of the time, a significant improvement over the 51.6% baseline of trusting all LLM outputs.

The anomaly detector achieves 89.3% recall, successfully identifying most claims that deviate from historical patterns. The 74.7% precision indicates some false positives where statistically unusual but correct predictions are flagged.

5.4 Case Study: Dementia Patient Search

A 72-year-old dementia patient was reported missing from a rural care facility. The terrain agent predicted "mountainous" (confidence: 0.85), while the status agent predicted "well" (confidence: 0.70).

Verification. For the terrain claim, single-hop lookup returns $P(\text{mountainous}|\text{dementia}) = 0.08$, triggering a LOW_PROB anomaly (severity 0.47). The claim is **FLAG**ged for human review; historical data shows dementia patients are found in mountainous terrain only 8% of the time (n=1,247 cases). For status, two-hop lookup returns $P(\text{well}|\text{dementia}, \text{rural}) = 0.62$, leading to **ACCEPT**. The patient was indeed found in flat terrain, validating the anomaly detection. Additional representative cases across all decision tiers are provided in Table 4 (Appendix).

5.5 Ablation Analysis

Anomaly Detection Impact. Figure 4 illustrates the contribution of each component. Comparing grounding-only

Method	Acc. Prec.	Coverage	Flag	Reject	Rej. Prec.	Ground.	Anom. F1
Raw LLM	51.6% ± 4.0%	100%	–	–	–	No	–
Rule-based (0.5–0.7)	51.6% ± 4.0%	100%	0%	0%	–	No	–
Rule-based (0.85)	52.3% ± 3.4%	51.8%	0%	48.2%	49.1%	No	–
Rule-based (0.90)	53.9% ± 5.8%	25.5%	0%	74.5%	49.2%	No	–
Self-Consistency†	51.1% ± 3.5%	96.6%	0%	3.4%	16.7%	No	–
Retrieval (RAG)	63.0% ± 3.8%	70.7%	0%	29.3%	82.6%	RAG	–
Grounding Only	65.3% ± 5.3%	54.5%	14.8%	10.2%	53.0%	KG	0.000
Full Pipeline	**71.0%** ± 4.7%	46.3%	14.5%	9.9%	58.8%	KG+RAG	**0.813**

†Self-Consistency is simulated: LLM confidence proxies consistency rate, avoiding 5× API cost.

Table 1. Method comparison on SAR verification task (n=100 scenarios, 5 seeds). Best results in bold.

Metric	Value
Accepted Precision	71.0% ± 4.7%
Reject Precision	58.8% ± 3.3%
Coverage (Accept Rate)	46.3% ± 2.4%
Anomaly Precision	74.7%
Anomaly Recall	89.3%
Anomaly F1	81.3%
Grounding Rate	70.7%
Two-hop Usage	11.7%
Single-hop Usage	59.0%

Table 2. Detailed performance breakdown (averaged over 5 seeds).

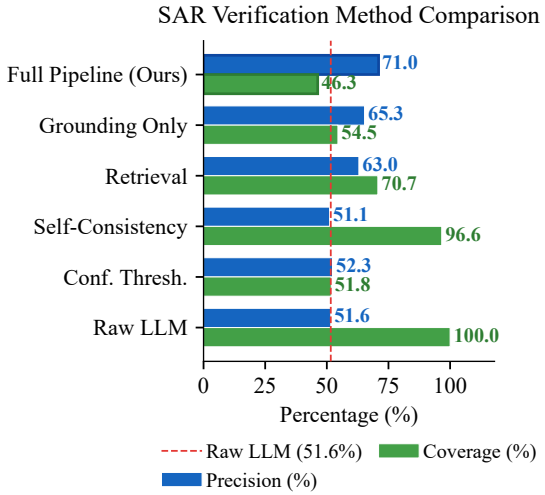

Figure 4. Accepted precision and coverage across methods.

(65.3% accepted precision) versus full pipeline (71.0%) isolates the contribution of anomaly detection: **+5.7%** absolute improvement. Without anomaly detection, the system cannot distinguish between well-supported and implausible claims among grounded predictions.

Two-Hop Reasoning. Two-hop reasoning (conditioning on subject + scenario) is used for 11.7% of grounded claims, providing more specific probability estimates for complex scenarios where single-hop statistics may be misleading.

KG vs. Simpler Verification Strategies. To assess whether the probabilistic KG structure is necessary, we compare against simpler alternatives. The Retrieval (RAG) baseline in Table 1 uses the same historical probability as our KG but applies a simple majority threshold (accept if $P \geq 0.50$) without anomaly detection or tiered decisions. It achieves 63.0% precision, above confidence-based methods but below our full pipeline (71.0%). A majority-vote strategy (always predict the most frequent terrain/status per subject category) would achieve high coverage but cannot distinguish between well-supported and implausible agent predictions, functioning as a static lookup rather than a verification mechanism. These comparisons demonstrate that the KG's value lies not in the probability estimates alone but in their integration with anomaly detection to identify statistically implausible claims.

Agent-Level Ablation. Ablating individual agent inputs (e.g., removing weather or medical features) requires re-running the full multi-agent pipeline with different configurations. Our experiments use simulated agent outputs to isolate the verification framework's contribution independently of upstream agent quality; end-to-end evaluation with selective agent removal is planned as future work.

Hyperparameter Sensitivity. We validate threshold choices via grid search over $\tau_{\text{low}} \in \{0.05\text{–}0.20\}$ and $\tau_{\text{mismatch}} \in \{0.25\text{–}0.40\}$. Accuracy ranges from 66.0% to 70.0% across configurations, with $(\tau_{\text{low}} = 0.15, \tau_{\text{mismatch}} = 0.40)$ achieving the best balance. For the minimum sample threshold $k \in \{20, 30, 40\}$, results are stable (65.7–65.8% accuracy), with $k = 30$ offering a good tradeoff between coverage (37 subject categories) and estimation reliability.

5.6 Error Analysis

We analyze the 31.6% incorrect decisions:

False Positives (10.5%): Predictions that aligned with population-level statistics but missed case-specific factors (67%) or fell within acceptable ranges but represented minority outcomes (33%). This reveals a fundamental limitation: population-level patterns cannot capture all case-specific nuances.

False Negatives (13.6%): Correct predictions that were

Decision	Count	Rate
Accept	175	43.8%
Reject	132	33.0%
Flag	36	9.0%
Candidate	57	14.2%

Table 3. Decision distribution across tiers (n=200 scenarios, 400 claims).

statistically unusual (e.g., a hiker actually found in urban terrain). The system is deliberately over-conservative, preferring false negatives (routing to human review) over false positives.

Grounding Gaps (29.3%): Claims where the subject category is absent from the KG receive CANDIDATE status.

5.7 Additional Analyses

Decision Distribution. Table 3 shows the system accepts 43.8% of claims, rejects 33.0%, and routes 23.2% to human review (flag + candidate), a conservative stance appropriate for safety-critical applications.

Performance by Claim Type. Terrain predictions achieve higher verified accuracy (71.2%) than status predictions (65.6%), likely because terrain distributions are more consistent while status outcomes depend on case-specific factors.

Computational Efficiency. Verification adds minimal overhead: KG construction takes 0.03 seconds, and 200 claims are verified in under 3 milliseconds (0.015 ms per claim).

6 Discussion and Limitations

When Verification Helps. Post-hoc verification is most valuable when LLM agents make overconfident predictions that deviate from historical patterns. The +19.4 percentage point improvement demonstrates that grounding in domain-specific statistical knowledge substantially enhances reliability. Ablation studies reveal that anomaly detection is the critical component; grounding alone is insufficient.

Verification Scope. Our evaluation focuses on terrain and status predictions that can be statistically grounded in ISRID. The framework is extensible to other ISRID fields (e.g., distance from IPP, search method), while real-time agent outputs (weather, photos) would require separate verification corpora.

Limitations.

- *Statistical vs. Absolute Correctness:* Our framework detects statistical anomalies, not causal errors. A prediction may be statistically unusual but correct (contributing to false negatives), or statistically typical but wrong for a specific case (contributing to false positives).
- *KG Coverage:* The system achieves 70.7% grounding rate; novel situations receive CANDIDATE status for human review.
- *Domain Specificity:* Generalization requires equivalent historical databases for new domains.
- *Agent Simulation:* Experiments use simulated agent outputs; future work will evaluate on live outputs from the deployed system.

Future Work. Key directions include: human-in-the-loop validation, online KG updates, semantic similarity for novel categories, contextual override combining population priors with case-specific evidence, and integration testing on real SAR operations.

7 Conclusion

We presented a post-hoc verification framework for multi-agent LLM systems in search and rescue decision support. By grounding agent outputs in a probabilistic knowledge graph derived from approximately 12,000 curated ISRID cases, our system detects statistical anomalies and produces tiered decisions with explicit evidence chains. The framework integrates with a seven-agent SAR system comprising weather, health, history, path analysis, photo analysis, interview, and knowledge fusion agents. Our verification scope focuses on terrain and status predictions, the claims most critical for search prioritization. Experiments demonstrate +19.4 percentage points improvement in accepted precision (71.0% vs. 51.6%) over unverified LLM outputs, with 81.3% anomaly detection F1. The framework provides a practical approach toward trustworthy AI in safety-critical emergency response domains. Our verification framework, aggregated probability tables (no individual case data), and evaluation scripts are publicly available.[1]

A Representative Verification Cases

Subject	Claim	Conf.	P(KG)	Decision	Outcome
ACCEPT: Prediction aligns with historical patterns					
Hiker	Terrain: Forest	0.78	0.67	ACCEPT	✓Correct
Child (1-3)	Status: Well	0.82	0.91	ACCEPT	✓Correct
FLAG: Requires human review					
Dementia	Terrain: Mountain	0.85	0.08	FLAG	✓Caught error
Hunter	Status: Injured	0.72	0.23	FLAG	× Correct but flagged
REJECT: Contradicts historical evidence					
Elderly	Terrain: Water	0.91	0.02	REJECT	✓Caught error
Hiker	Status: Deceased	0.88	0.04	REJECT	✓Caught error
CANDIDATE: Insufficient KG coverage					
Climber	Terrain: Mountain	0.75	–	CANDIDATE	Unverified

Table 4. Representative verification cases across decision tiers.

Ethics Statement

Data Privacy. The ISRID database contains de-identified historical SAR records. We publish only aggregated probability distributions; no individual case data is shared. All test scenarios are synthetically generated.

Responsible AI. Our framework is designed as a decision *support* tool, not an autonomous decision-maker. Human SAR coordinators must review all recommendations, and the system provides evidence chains to enable informed judgment. The framework complements, rather than replaces, human expertise.

[1] https://github.com/yuki011121/sar-verification-framework

Acknowledgments

We gratefully acknowledge Robert J. Koester for creating and maintaining the International Search and Rescue Incident Database (ISRID) and for his seminal work "Lost Person Behavior: A Search and Rescue Guide on Where to Look—for Land, Air and Water," which provided the foundational data for our probabilistic knowledge graph. We thank the SAR community for their contributions to ISRID over decades of data collection.

References

Belden, A. 2024. *Causal Inference Using Bayesian Network For Search And Rescue*. Master's thesis, California Polytechnic State University. Available at https://digitalcommons.calpoly.edu/theses/2797.

Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; and Weston, J. 2023. Chain-of-Verification Reduces Hallucination in Large Language Models. *arXiv preprint arXiv:2309.11495*.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325*.

Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; and Larson, J. 2024. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130*.

Gao, L.; Dai, Z.; Pasupat, P.; Chen, A.; Chaganty, A. T.; Fan, Y.; Zhao, V. Y.; Lao, N.; Lee, H.; Juan, D.-C.; et al. 2023. RARR: Researching and Revising What Language Models Say, Using Language Models. *arXiv preprint arXiv:2210.08726*.

Garcez, A. d.; and Lamb, L. C. 2023. Neurosymbolic AI: The 3rd Wave. *Artificial Intelligence Review*, 56(11): 12387–12406.

Hashimoto, A.; Heintzman, L.; Koester, R. J.; and Abaid, N. 2022. An agent-based model reveals lost person behavior based on data from wilderness search and rescue. *Scientific Reports*, 12(1): 5873.

Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; Melo, G. d.; Gutierrez, C.; Kirrane, S.; Gayo, J. E. L.; Navigli, R.; Neumaier, S.; et al. 2021. Knowledge Graphs. *ACM Computing Surveys*, 54(4): 1–37.

Hong, S.; Zhuge, M.; Chen, J.; Zheng, X.; Cheng, Y.; Zhang, C.; Wang, J.; Wang, Z.; Yau, S. K. S.; Lin, Z.; et al. 2023. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. *arXiv preprint arXiv:2308.00352*.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12): 1–38.

Jiang, J.; Zhou, K.; Dong, Z.; Ye, K.; Zhao, W. X.; and Wen, J.-R. 2023. StructGPT: A General Framework for Large Language Model to Reason over Structured Data. *arXiv preprint arXiv:2305.09645*.

Kim, B. 2025. *Opening the Black Box with REGAL: A Novel Explainable AI Approach to Uncover Key Predictors in Search and Rescue Success*. Master's thesis, California Polytechnic State University. Available at https://digitalcommons.calpoly.edu/theses/3029.

Koester, R. J. 2008. *Lost Person Behavior: A Search and Rescue Guide on Where to Look—for Land, Air and Water*. dbS Productions.

Koester, R. J. 2023. International Search and Rescue Incident Database (ISRID). https://www.dbs-sar.com/SAR_Research/ISRID.htm. Accessed: 2024.

Kurfess, F.; Tan, Y.; and Delgado, E. 2025. AI-Powered Search and Rescue: An Agentic System Utilizing Knowledge Graphs and LLMs. In *NxtAI 2025*. San Francisco, CA, USA.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33: 9459–9474.

Li, G.; Hammoud, H. A. A. K.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. *Advances in Neural Information Processing Systems*, 36.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; et al. 2023. AgentBench: Evaluating LLMs as Agents. *arXiv preprint arXiv:2308.03688*.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *arXiv preprint arXiv:2305.14251*.

Nguyen, J.; Joseph, C.; Richardson, B.; Hayes, R.; Pakula, R.; and Koester, R. J. 2023. Finding a Needle in the Haystack: Predicting the Location of Lost People Using Agent-Based Modeling and Behavioral Inertia. In *2023 Systems and Information Engineering Design Symposium (SIEDS)*, 78–83.

Pan, S.; Luo, L.; Wang, Y.; Chen, C.; Wang, J.; and Wu, X. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 3580–3599.

Queralta, J. P.; Taipalmaa, J.; Pullinen, B. C.; Sarker, V. K.; Gia, T. N.; Tenhunen, H.; Gabbouj, M.; Raitoharju, J.; and Westerlund, T. 2020. Collaborative Multi-Robot Search and Rescue: Planning, Coordination, Perception, and Active Vision. *IEEE Access*, 8: 191617–191643.

Shaabani, E.; Alvari, H.; Shakarian, P.; and Snyder, J. E. K. 2016. MIST: Missing Person Intelligence Synthesis Toolkit. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM)*, 1843–1867.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171*.

Washington, C. 2018. *Topographic Maps: Image Processing and Path-Finding*. Master's thesis, California Polytechnic State University. Available at https://digitalcommons.calpoly.edu/theses/2504.

Wei, J.; Yang, C.; Song, X.; Lu, Y.; Hu, N.; Huang, J.; Tran, D.; Peng, D.; Liu, R.; Huang, D.; et al. 2024. Long-form Factuality in Large Language Models. *arXiv preprint arXiv:2403.18802*.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155*.

Young, C. S. 2022. *Intelligent Search: Managing the Intelligence Process in the Search for Missing Persons*. Charlottesville, Virginia: dbS Productions LLC. ISBN 978-1-879471-62-7.

Young, C. S.; Kurfess, F.; and Bloom, G. 2024. The Search Intelligence Process Using Artificial Intelligence. *Journal of Search and Rescue*, 7.

Young, C. S.; Kurfess, F.; and Bloom, G. 2025. AI4SAR: Enhancing Missing Person Searches with Advanced Data Analysis and Artificial Intelligence. In *National Missing and Unidentified Persons Conference 2025*.