how to judge a teams performance where it has been given an AI based challenege e.g making sense of a lot of logs in a SOC environment

To judge a team evaluating AI in a Security Operations Center (SOC) context—specifically for "making sense of logs"—you need to move beyond simple accuracy scores. The challenge in SOC is not just finding the threat (detection) but explaining it clearly enough for an analyst to act (contextualization) without overwhelming them (alert fatigue).

The following framework breaks evaluation into five weighted pillars. You can use this as a rubric or scorecard for your challenge.

Executive Summary: The 5 Pillars of Evaluation

Pillar	Weight	Core Question	Key Metric
1. Efficacy & Accuracy	30%	Did it find the real threats without crying wolf?	Precision/Recall, False Positive Rate
2. Interpretability	25%	Does the AI explain *why* something is a threat?	Narrative Quality, Evidence Linking
3. Operational Utility	20%	Does this actually save an analyst time?	Click-reduction, Actionability
4. Log "Sense-making"	15%	Can it handle volume, noise, and formatting?	Correlation capabilities, Noise filtering
5. Innovation & Build	10%	Is the technical approach robust and novel?	Latency, Scalability, Agent Architecture

Detailed Judging Criteria

1. Efficacy & Accuracy (30%)

Focus: The mathematical performance of the model.

- **False Positive Rate (FPR):** This is the most critical metric in SOC. If an AI flags normal admin activity as malicious, it creates "alert fatigue." Teams should be penalized heavily for high FPR.
- **Recall (Detection Rate):** Did the team find the "hidden needles" in your log dataset?

    - *Test:* Inject known attack patterns (e.g., Password Spraying, Data Exfiltration) and count how many were flagged.
- **Entity Extraction:** How accurately did the AI extract key entities (IPs, Usernames, File Hashes)?
    - *Failure Mode:* Hallucinating an IP address that doesn't exist in the logs.

2. Interpretability & "Why" (25%)

Focus: The "Black Box" problem. In security, an alert without a reason is useless.

- **Narrative Generation:** Can the AI summarize *what happened* in plain English?
    - *Good:* "User ADMIN_01 logged in from an unusual geo-location (China) at 3 AM and immediately downloaded 5GB of data."
    - *Bad:* "Anomaly detected in Auth Logs. Score 0.98."
- **Evidence Linking:** Does the AI cite the specific log lines that led to its conclusion? (e.g., "See log entries #450–#460").
- **Chain of Thought:** specifically for GenAI/LLMs, does the system show its reasoning steps? (e.g., "First I saw a login failure, then a success, then a privilege escalation.")

3. Operational Utility (20%)

Focus: The User Experience (UX) for the Tier-1 Analyst.

- **Actionability:** Does the AI suggest a valid next step? (e.g., "Isolate Host," "Reset Password," vs. just "Investigate").
- **Context Enrichment:** Did the AI look up external info to help make sense of the logs?
    - *Example:* If a log shows a connection to IP `1.2.3.4`, did the AI auto-check if that IP is a known malicious server (Threat Intel)?
- **Triage Categorization:** Did it correctly classify the severity? (Critical vs. Informational).

4. Log "Sense-making" Capabilities (15%)

Focus: Handling the messy reality of logs.

- **Correlation:** Can the team link logs across different sources?
    - *Challenge:* Connect a *Network Log* (firewall allow) with an *Endpoint Log* (process start) to show a complete attack chain.
- **Noise Reduction:** Did the solution successfully group repetitive logs? (e.g., "Compressed 5,000 failed login attempts into 1 alert").
- **Unstructured Parsing:** How well did it handle non-standard or messy log formats without rigid regex rules?

5. Innovation & Technical Architecture (10%)

- **Latency/Speed:** Did it process the logs in near real-time, or did it take 10 minutes to analyze a batch?
- **Cost Efficiency:** (If applicable) Did they use a massive, expensive model (e.g., GPT-4o) for everything, or a smaller, efficient model (e.g., Llama-3-8B) for bulk processing?
- **Security of the AI:** Did they implement guardrails to prevent Prompt Injection? (e.g., Can a malicious log entry trick the AI into saying "System Secure"?)

How to Run the Evaluation (The "Test Set")

To judge fairly, you need a **Golden Dataset**—a set of logs where *you* know the ground truth.

1. **The "Clean" Set:** 10,000 lines of normal business traffic (people logging in, printing, browsing).
   - *Goal:* Test False Positives. The AI should be quiet here.
2. **The "Attack" Set:** The Clean Set + 3 specific attack scenarios injected (e.g., SQL Injection, Brute Force, Malware Beaconing).
   - *Goal:* Test Recall. Did it find all 3?
3. **The "Confusing" Set:** Logs that *look* bad but are benign (e.g., a system admin running a scanner).
   - *Goal:* Test Reasoning. The AI should recognize this is authorized behavior, or at least flag it as "Suspicious but likely Admin."

Sample Scorecard for Judges

Team Name: _____	Score (1-10)	Notes/Observations
Accuracy: Did they find the 3 hidden attacks?	___/10	*Found 2/3 attacks. 1 False Positive.*
Clarity: Is the AI summary easy to read?	___/10	*Summary was vague; didn't explain 'why'.*
Evidence: Do they link back to raw logs?	___/10	*Yes, clickable links to raw log lines.*
Correlation: Did they link Network + Identity?	___/10	*Failed to link IP to Username.*
UI/UX: Is the dashboard intuitive?	___/10	*Clean interface, good visualization.*
TOTAL	**___/50**	

⁂

1. https://radiantsecurity.ai/learn/evaluate-ai-soc-analysts/
2. https://smallestai.thescrs.org/judging-criteria/
3. https://www.dropzone.ai/blog/how-to-evaluate-an-ai-soc-analyst
4. https://dl.acm.org/doi/10.1145/3755881.3755888
5. https://unstop.com/blog/how-to-judge-a-hackathon

6. https://www.cybersecuritytribe.com/articles/4-key-metrics-ai-socs-will-impact
7. https://arxiv.org/html/2502.00677v1
8. https://apartresearch.com/sprints/agent-security-hackathon-2024-10-04-to-2024-10-07
9. https://underdefense.com/blog/soc-metrics/
10. https://ieeexplore.ieee.org/iel8/6287639/10820123/11141466.pdf
11. https://www.kaggle.com/competitions/rmit-gen-ai-and-cyber-security-hackathon
12. https://www.conifers.ai/blog/soc-metrics-kpis-how-to-measure-ai-soc-performance
13. https://www.prophetsecurity.ai/blog/soc-metrics-that-matter-mttr-mtti-false-negatives-and-more
14. https://arxiv.org/html/2507.02390v1
15. https://ai-bias-bounty-hackathon.devpost.com
16. https://blog.7ai.com/6-agentic-ai-soc-metrics-measuring-the-value-of-ai-agents-in-security-operations
17. https://www.sciencedirect.com/science/article/abs/pii/S0167404824003213
18. https://taikai.network/en/blog/hackathon-judging
19. https://coralogix.com/ai-blog/evaluation-metrics-for-ai-observability/
20. https://github.com/Jorge-Tejero-Fdez/ThreatLogLLM/