"The question of whether machines can think is about as relevant as the question of whether submarines can swim."

Dijkstra, E. W. (1984). The threats to computing science. In Selected Writings on Computing: A Personal Perspective. Springer, New York.

16/08/2025

**Assessment 2**:  
**Scoping a research problem Report**

**Project Title:   
Benchmarking Generative AI Token Use in Cybersecurity Operations**

**Name:  
Mohamed Zeyada - 11693860**

**Supervisor:  
Dr. Gowri Ramachandran**

**Cluster: 8**

# 1. Research Background & Literature Analysis

Generative Artificial Intelligence (GenAI) is transforming cybersecurity operations, especially within Security Operations Centers (SOC) and Governance, Risk, and Compliance (GRC) workflows. Large Language Models (LLMs) such as GPT-4, Claude, and Gemini are now used for high-stakes tasks including incident report generation, threat intelligence summarizations, and compliance documentation. These applications can improve analyst efficiency and analytic scope, but in security-critical environments, accuracy, compliance adherence, and cost efficiency are non-negotiable. SOC analysts must triage incidents within minutes, while GRC professionals work to strict audit deadline verbose or inefficient AI outputs can hinder operational resilience and compliance readiness.

A key challenge is the **cost–performance trade-off**: richer prompts may improve completeness but consume more tokens, increasing API costs and—per evidence on verbosity bias—over-reward long responses relative to factual quality. In SOC/GRC contexts, conciseness is vital: verbose AI-generated reports may score highly yet delay decision-making.

## Literature Review & Integration

CySecBench (Wahréus, Hussain, & Papadimitratos, 2025)

**Contribution:** Provides a cybersecurity benchmark whose rubric is adapted in this project into seven SOC/GRC-relevant dimensions:

* Technical Accuracy
* Actionability
* Completeness
* Compliance Alignment
* Risk Awareness
* Relevance
* Clarity

Includes a large dataset of domain-specific prompts across multiple attack categories, which can be selectively sampled for evaluation.

**Gap:** The dataset is static and does not address:

* Token-cost analytics
* Adaptive SOC/GRC tasks
* Evolving regulatory requirements

**Integration:** Serves as a baseline rubric and dataset reference. This project narrows the dataset to prompts relevant to:

* GRC
* Incident Response
* SOC contexts

It is supplemented with adaptive prompts derived from:

* SOC workflows
* Compliance policies
* Live cyber threat intelligence (CTI)

DefenderBench (Zhang et al., 2025)  
  
**Contribution:** Supports multi-task evaluation under SOC-like conditions, including:

* Malicious content detection
* Vulnerability remediation
* Cybersecurity question answering (QA)

**Gap:** Does not include:

* SOC/GRC rubric mapping
* Cost–performance tracking

**Integration:** Applies its task diversity and agentic scaffolding in combination with:

* SOC/GRC rubric dimensions
* Token-cost and performance analytics

Same Evaluation, More Tokens (Domhan & Zhu, 2025)  
**Contribution:** Identifies verbosity bias in prompt-based evaluations and introduces **Focus Sentence Prompting (FSP)** to normalize prompt length and reduce linguistic inflation.

**Gap:** Not tailored to SOC/GRC contexts. Requires:

* Domain-specific mapping of FSP
* Bias-adjusted scoring aligned with SOC/GRC rubric dimensions

**Integration:** Implements FSP as the default bias-mitigation module within this project’s evaluation framework.

Chroma Generative Benchmarking (Hong et al., 2025)  
**Contribution:** Generates representative queries from seed documents and validates them using:

* Large Language Model (LLM) judges
* Distributional checks such as **Kullback–Leibler (KL) divergence**

**Gap:** Not specialized for:

* Cybersecurity compliance contexts
* CTI-driven (Cyber Threat Intelligence) prompt drift

**Integration:** Adapts the query generation framework to support:

* SOC/GRC policy updates
* Live CTI feed integration for dynamic prompt evaluation

## Environmental Scan

Commercial SOC/GRC AI tools (e.g., Microsoft Security Copilot, Recorded Future AI Summaries, CrowdStrike Charlotte AI) overlap in incident summarization, feed parsing, and compliance drafting, but they:

1. Lack transparent rubric-based scoring for quality assurance.
2. Do not reveal token cost trade-offs for budget optimization.
3. Use static or opaque datasets, missing evolving threats.
4. Provide proprietary, non-reproducible evaluation pipelines.

No current tool integrates rubric evaluation, token analytics, bias mitigation, and adaptive dataset generation in an open, reproducible framework.

### Identified Gap

No platform currently offers a reproducible SOC/GRC-aligned LLM benchmarking framework that:

1. Uses SOC/GRC-relevant rubric scoring.
2. Tracks token-level cost with trade-off visualization.
3. Mitigates verbosity bias.
4. Adapts benchmarks to emerging threats and compliance changes.

### Proposed Solution

. We will build a **cost–quality-aware SOC/GRC benchmarking platform** that:

* **Benchmarks LLMs** by linking token usage and API costs to output quality across seven SOC/GRC-relevant dimensions: *Technical Accuracy, Actionability, Completeness, Compliance Alignment, Risk Awareness, Relevance, and Clarity.*
* **Applies Focus Sentence Prompting (FSP)** by default to mitigate verbosity bias, ensuring fair cost–quality comparison across models and prompt lengths.
* **Uses adaptive prompts** inspired by Chroma but tailored to SOC workflows, compliance policies, and live CTI, so evaluations remain realistic and current.
* **Optionally incorporates DefenderBench task formats** to enrich SOC/GRC evaluation diversity.
* **Delivers interactive dashboards** for cost vs. quality insights, helping practitioners identify optimal trade-offs for SOC and GRC operations.

### Novelty

The novelty lies in **explicitly linking token cost analysis with multidimensional quality evaluation** in SOC/GRC contexts, while ensuring fairness via bias mitigation and adaptability via evolving prompt sets. This builds **reproducible, real-world baselines** for cost-effective and responsible use of generative AI in cybersecurity.

## Literature to Feature Mapping

|  |  |  |  |
| --- | --- | --- | --- |
| Source | Contribution | Gap | Feature Integration |
| Wahréus, Hussain, & Papadimitratos (2025) – *CySecBench* | Provides a cybersecurity-focused benchmark and rubric used as a foundation for this project’s adapted SOC/GRC-relevant rubric | Dataset is static; lacks token-cost tracking | Use rubric as baseline, extend with cost analytics |
| Zhang et al. (2025) – *DefenderBench* | Multi-task cybersecurity evaluation under SOC-like conditions | No SOC/GRC rubric alignment; no token-cost tracking | Integrate diverse tasks into SOC/GRC framework with cost monitoring |
| Domhan & Zhu (2025) – *Same Evaluation, More Tokens* | Identifies verbosity bias; introduces Focus Sentence Prompting (FSP) | Not domain-specific; lacks SOC/GRC adaptation | Apply FSP as default bias-mitigation |
| Hong et al. (2025) – *Chroma Benchmarking* | Adaptive prompt generation; validates representativeness with LLM judges | Not tailored to cybersecurity compliance or CTI-driven drift | Adapt to SOC/GRC policy updates and live CTI feeds |

# 2. Research Problem Statement

What**:**  
Current cybersecurity LLM benchmarking is fragmented. Some benchmarks use SOC/GRC rubrics but lack cost analysis (CySecBench), others offer task realism without SOC/GRC scoring (DefenderBench), and some adapt datasets but without security focus (Chroma). No unified framework combines all.

Why:  
SOC and GRC teams face:  
• Tight budgets—API usage directly impacts cost.  
• Strict compliance—outputs must meet regulatory standards.  
• Rapidly changing threats—static benchmarks miss emerging risks.

Without integrated cost–performance benchmarking, teams risk overspending, using suboptimal prompts, and relying on outdated datasets.

Novelty:  
This project will be the first continuous SOC/GRC benchmarking platform that:

1. Combines CySecBench rubrics with DefenderBench realism.
2. Integrates token cost analytics with model pricing.
3. Uses FSP by default to counter length bias.
4. Generates adaptive benchmarks from SOC/GRC policies and CTI feeds.

## Objectives (End of Semester):

1. Web-based SOC/GRC benchmarking tool with dataset selection, evaluation, and analytics.
2. Cost–quality trade-off dashboards.
3. Reproducible pipelines with fixed seeds and dataset versioning.
4. Live threat-adaptive benchmark generation from policy + CTI.

Expected Outputs:  
a. SOC/GRC benchmark datasets (static + adaptive).  
b. Token usage & cost analytics dashboard.  
c. Bias-mitigated scoring pipeline (raw vs. FSP-adjusted).  
d. public methodology and evaluation reports.

# 3. Research Questions

RQ1. How does prompt length influence LLM output quality and cost efficiency in SOC/GRC tasks?  
***What****:* Quantify the link between prompt length, quality scores, and token/API costs for SOC/GRC tasks (e.g., incident reporting, compliance mapping, threat intel summaries).  
***Why****:* Prompt design impacts quality and cost; knowing the point where quality gains plateau enables optimal deployment.  
***How****:*  
• Select CySecBench-style tasks.  
• Craft short, medium, and long prompt variants.  
• Score with a seven-dimension adapted rubric.  
• Log tokens and costs.  
• Apply FSP for bias-corrected scores.  
***Contribution****:* Evidence-based prompt engineering for SOC/GRC.  
*Outputs:* Cost–quality trade-off curves; recommended context windows; prompt design playbook.  
*Feasibility:* Uses public datasets, standard LLM APIs, and existing compute resources.

RQ2. Can adaptive generative benchmarking from SOC/GRC documents and CTI improve evaluation coverage and relevance over static datasets?  
**What***:* Test if a Chroma-inspired pipeline produces benchmark items better aligned with current SOC/GRC needs.  
**Why***:* Static benchmarks drift; regulatory and threat updates must be reflected.  
***How****:*  
• Ingest SOC/GRC policies and CTI.  
• Generate candidate prompts via LLM.  
• Filter with relevance judge.  
• Compare coverage and performance vs. CySecBench baselines using KL divergence.  
*Contribution:* Validates adaptive benchmarking for long-term relevance.  
*Outputs:* Versioned adaptive benchmark sets; comparative performance analyses.  
*Feasibility:* Uses public policy docs, open CTI feeds, and adapted generative scripts.

# References

* Catchpoint. (2024). *GenAI performance benchmarking*. Catchpoint. <https://www.catchpoint.com/learn/gen-ai-benchmark>
* Domhan, T., & Zhu, D. (2025). *Same evaluation, more tokens: On the effect of input length for machine translation evaluation using large language models*. arXiv. <https://arxiv.org/abs/2505.01761>
* Hong, K., Troynikov, A., Huber, J., & McGuire, M. (2025). *Generative benchmarking*. Chroma. <https://research.trychroma.com/generative-benchmarking>
* Srivastava, A., et al. (2023). *Beyond the imitation game: Quantifying and extrapolating the capabilities of language models*. *Transactions on Machine Learning Research*. <https://openreview.net/forum?id=uyTL5Bvosj>
* Wahréus, J., Hussain, A. M., & Papadimitratos, P. (2025). *CySecBench: Generative AI-based cybersecurity-focused prompt dataset for benchmarking large language models*. arXiv. <https://arxiv.org/abs/2501.01335>
* Zhang, C., Côté, M.-A., Albada, M., Sankaran, A., Stokes, J. W., Wang, T., Abdi, A., Blum, W., & Abdul-Mageed, M. (2025). *DefenderBench: A toolkit for evaluating language agents in cybersecurity environments*. arXiv. <https://arxiv.org/abs/2506.00739>

# Appendix

This appendix provides a structured evaluation of three key papers that directly inform the design, methodology, and scope of the proposed SOC/GRC-specific LLM benchmarking platform. Each evaluation follows the required format: source information, research objectives, problems and gaps addressed, main findings, limitations, implications, and the specific gap our research will address.

## 1. CySecBench (Wahréus, Hussain, & Papadimitratos, 2025)

**Source Information**:  
IEEE/Arxiv preprint; cybersecurity benchmark for SOC/GRC LLM evaluation.

**Research Questions/Objectives**:  
To develop a reproducible, rubric-based benchmarking framework for evaluating LLM performance in cybersecurity operations, using a domain-specific dataset and clear scoring dimensions.

**Problems and Gaps Addressed**:  
Lack of SOC/GRC-focused, rubric-based benchmarks with realistic domain prompts.

**Main Findings and Conclusions**:  
Introduces a six-metric rubric and a dataset of 12,662 prompts across ten attack categories (e.g., malware, intrusion, cloud). Demonstrates feasibility of reproducible scoring in SOC-like tasks.

**Limitations and Weaknesses:**

* Dataset is static
* No adaptive updates for emerging threats or regulatory changes
* No token-cost analytics

**Implications and Suggestions**:  
Provides a strong baseline for reproducible SOC/GRC evaluation. Needs integration with cost-performance analysis and adaptive dataset updates.

**How Our Research Fills the Gap**:  
Adopts and extends CySecBench by mapping evaluation to our SOC/GRC-specific seven-dimension rubric (Technical Accuracy, Actionability, Completeness, Compliance Alignment, Risk Awareness, Relevance, Clarity). Adds token-cost analytics and incorporates adaptive prompt generation from SOC/GRC policies and live CTI feeds.

## 2. DefenderBench (Zhang et al., 2025)

**Source Information**:  
Arxiv preprint; multi-task evaluation toolkit for cybersecurity LLM agents.

**Research Questions/Objectives** To evaluate LLMs on a range of SOC-like tasks including:

* Malicious content detection
* Vulnerability remediation
* Security Q&A under realistic operational constraints

**Problems and Gaps Addressed** Existing cybersecurity LLM evaluations focus on narrow task sets or lack operational realism.

**Main Findings and Conclusions** Demonstrates broad task coverage and agentic evaluation workflows for SOC environments.

**Limitations and Weaknesses**

* No SOC/GRC-specific rubric alignment
* Lacks token usage and cost analytics

**Implications and Suggestions** Can be adapted to our SOC/GRC rubric to maintain realism while integrating cost analysis.

**How Our Research Fills the Gap** Incorporates DefenderBench's diverse task formats but scores them using our SOC/GRC seven-dimension rubric. Logs token consumption and visualizes cost–performance trade-offs.

## 3. Same Evaluation, More Tokens (Domhan & Zhu, 2025)

**Source Information**:  
Arxiv preprint; Study on verbosity bias in LLM evaluation and mitigation strategies.

**Research Questions/Objectives**:  
To investigate the impact of verbosity bias in LLM evaluations and propose mitigation via Focus Sentence Prompting (FSP).

**Problems and Gaps Addressed**:  
Longer responses tend to receive higher scores irrespective of quality, distorting evaluation fairness.

**Main Findings and Conclusions**:  
Quantifies verbosity bias and shows that FSP normalises prompt length without sacrificing essential content. *This ensures that cost–quality analysis reflects true content quality rather than inflated verbosity.*

**Limitations and Weakness**:

Not tailored to SOC/GRC tasks; requires domain adaptation of FSP and bias-adjusted scoring.

**Implications and Suggestions**:

Integrating FSP into domain-specific evaluation pipelines can improve fairness.

**How Our Research Fills the Gap**:

Embeds FSP into our SOC/GRC evaluation framework to ensure that rubric scores reflect content quality rather than length, applied alongside our seven-dimension rubric.

## SOC/GRC-Specific Evaluation Rubric

The evaluation framework applies the following seven-dimension rubric to all benchmarked outputs:

|  |  |  |
| --- | --- | --- |
| Dimension | What It Evaluates | Scoring Guide |
| 1. Technical Accuracy | Factual correctness, terminology usage, and proper alignment with real-world cybersecurity models | Low: Hallucinated threats/tools; misused concepts (e.g., hashing vs encryption). High: Verifies claims, uses MITRE, CVEs, or NIST references precisely |
| 2. Actionability | Does the response deliver step-by-step, operationally usable security guidance? | Low: Generic, abstract summaries without task guidance. High: Detailed instructions (e.g., containment, triage steps, ISO/NIST controls) |
| 3. Completeness | Fulfillment of all prompt components and security-relevant context coverage | Low: Omits important prompt tasks or partial info. High: Every aspect of the prompt is addressed with contextual detail |
| 4. Compliance Alignment | Whether recommendations adhere to regulatory frameworks and policies | Low: Gives illegal or policy-violating suggestions. High: Explicitly maps advice to NIST 800-53, ISO 27001, GDPR, etc. |
| 5. Risk Awareness | Identifying, analyzing, and mitigating risks in the context of the response | Low: Ignores threat implications, impact, or risk likelihood. High: Includes threat modeling, risk matrixing, or control strategies |
| 6. Relevance | Alignment of response content with the specific context and goal of the prompt | Low: Response drifts from prompt intent or introduces unrelated content. High: Directly addresses task goals using domain-relevant terminology and scenario framing |
| 7. Clarity | Linguistic and structural clarity of the response; ease of comprehension for practitioners | Low: Ambiguous, jargon-heavy, or convoluted language impedes understanding. High: Clear, concise structure using readable language; includes bulleting or step formatting where appropriate |