1. Executive Summary

As federal agencies navigate increasing cybersecurity demands alongside growing expectations to adopt artificial intelligence (AI), the question is no longer *whether* AI will be used, but *how* it can be used responsibly, securely, and effectively. Within governance, risk, and compliance (GRC) functions, AI-enabled tools offer opportunities to improve speed, scale, and consistency across cybersecurity programs, while also raising important questions about trust, transparency, and decision authority.

To better understand how AI performs in real Risk Management Framework (RMF) assessment contexts, NR Labs conducted a hands-on evaluation of several large language models (LLMs) as decision-support tools for assessment and monitoring activities. Using a consistent subset of controls, assessment objectives, prompts, and evidentiary artifacts, we examined model outputs against human assessor expectations across three focus areas: accuracy, gap detection, and reasoning quality.

Our testing revealed that AI models can meaningfully support assessors by summarizing documentation, highlighting gaps, and offering explanations that support human decision-making. However, performance varied in the consistency with which models applied judgment, evaluated partial evidence, and articulated conclusions. The quality of evidence was a critical factor influencing outcomes, and no model consistently replicated human assessor reasoning across all evaluated scenarios.

2. Background and Purpose

As agencies manage growing system inventories, continuous monitoring requirements, and persistent resource constraints, AI-enabled tools are increasingly viewed as potential force multipliers. Within GRC functions, these tools offer opportunities to accelerate documentation review, improve consistency in analysis, and provide more timely insight into organizational risk.

Despite this interest, comparatively little attention has been given to how AI models perform at the control and evidence level, where assessor judgment, context, and nuance remain essential. Much of the current discussion emphasizes theoretical benefits or high-level automation, while far less attention is given to how these tools function in realistic assessment scenarios.

Without structured evaluation, organizations may overestimate model capabilities or place undue trust in outputs that appear authoritative but lack sufficient supporting evidence. In RMF assessments, where conclusions directly inform authorization decisions and ongoing risk posture, such misjudgments can have meaningful operational and compliance consequences.

This paper builds on applied testing to help close this gap and provide practical insight into the use of AI language models in RMF assessment activities.

3. Scope of the Evaluation

This evaluation examined how multiple AI language models perform when used as decision-support tools for discrete RMF assessment activities. The scope focused on model behavior during common assessor tasks, rather than on producing generalized model rankings or measuring broad, general-purpose AI capability.

Specifically, the evaluation examined:

- The use of AI language models to support RMF-aligned assessment and monitoring activities
- Model behavior when interpreting control requirements and assessment objectives
- Evaluation of supporting artifacts with varying levels of quality and completeness
- Identification of control gaps and articulation of assessment reasoning
- Performance patterns across accuracy, gap detection, and reasoning quality

To enable meaningful comparison, testing was conducted using a consistent set of prompts, assessment objectives, and evidentiary artifacts across all models.

4. Controls, Objectives, and Artifact Context

Controls Evaluated

The evaluation focused on a targeted set of NIST SP 800-53 Revision 5 controls selected to represent different phases of the RMF and varying types of assessor judgment. These controls were chosen because they require interpretation of documentation quality, procedural rigor, and supporting evidence, and supporting evidence in various formats such as policy documents, spreadsheets, and screenshots of system logs.

Controls Evaluated	Control Focus
PL-2	System security & privacy planning documentation
AU-6	Audit review, analysis, and reporting
SR-6	Supplier assessments & supply chain risk

RMF Phase Alignment

The following illustrates the primary RMF phases supported by each control. Highlighted phases indicate the primary RMF activities supported.

Control	Primary RMF Phase (s)
PL-2	Select
AU-6	Monitor
SR-6	Select, Implement, Monitor

Note: Primary phases reflect the RMF activities most directly supported by each control within this evaluation.

Assessment Objectives

For each control, testing was organized around defined assessment objectives aligned to common RMF evaluation activities. These objectives were designed to mirror how human assessors review control implementation and supporting evidence during formal assessments.

These objectives focused on the model's ability to:

- Review documented control implementation statements
- Evaluate procedures and processes supporting the control
- Assess the sufficiency and relevance of evidence artifacts
- Identify gaps or weaknesses relative to control requirements
- Explain assessment conclusions in clear, defensible terms

Each objective was evaluated using standardized prompt types to ensure consistency across models and controls.

Artifact Types and Conditions

Each control was evaluated using artifacts designed to reflect realistic RMF assessment inputs. Artifacts were developed by assessors to mirror documentation and evidence commonly encountered during formal RMF assessments, while avoiding the use of system-specific or sensitive information.

Artifact types were standardized within each control to ensure consistent evaluation across large language models (LLMs).

Artifact Types Included

- **Implementation statements -** descriptions of how the control is implemented
- **Documented procedures -** processes and workflows supporting the control
- **Evidence artifacts -** materials showing control operation or enforcement

To evaluate model sensitivity to evidence quality, each artifact displayed two versions, a weak and strong variant.

- **Weak artifacts** - incomplete, vague, inconsistently applied documentation or evidence
- **Strong artifacts** - clearly defined, procedurally mature documentation or evidence aligned with expected compliance standards

For a given control, the same artifacts were evaluated across all LLMs under both conditions to enable consistent comparison of results and isolate performance differences attributable to model behavior rather than input variation.

5. Methodology

Test Design

The evaluation followed a standardized test design to ensure that observed differences in model behavior reflected model characteristics rather than differences in inputs or configuration.

Key elements of the test design included:

- Control-specific assessment objectives applied consistently across all models evaluated for each control
- Control-specific artifacts applied consistently across all models
- Prompts and assessment parameters were standardized for consistent cross-model comparison
- Independent evaluation of weak and strong artifact conditions
- No prompt tuning or model-specific optimization
- Human assessor judgment used as the reference point for expected outcomes

Assessment results were documented using standardized evaluation workbooks. Separate workbooks were maintained for each control, with individual worksheets used to capture model outputs and assessor observations for each model evaluated.

Models Evaluated

The evaluation included the following commercially available models, tested without prompt tuning or model-specific optimization.

Provider	Model
Open-AI	GPT –4 Gov
Open-AI	GPT 4o-mini Gov
Anthropic	Claude 3.7 Sonnet Gov
Google	Gemini 2.5 Pro
Microsoft	Copilot 5.1
Meta	Llama 3.3

Evaluation Focus Areas

Model performance was assessed across three primary focus areas selected to reflect core RMF assessment activities.

- **Accuracy** - The degree to which model outputs aligned with control requirements and expected assessor determinations based on the provided artifacts**.**
- **Gap Detection** - The model's ability to identify missing, incomplete, or insufficient elements relative to control requirements and assessment objectives.
- **Reasoning Quality** - The clarity, coherence, and defensibility of the model's explanations supporting its assessment conclusions.

Scoring Approach

Model outputs were evaluated using a qualitative scoring scale applied consistently across all focus areas. Scores reflected alignment with human assessor judgment, including both the assessment outcome and the clarity and defensibility of the reasoning provided.

Scoring was intentionally kept high-level to identify patterns in model behavior rather than to produce precise numerical rankings or designate a single “best” model.

Consistency and Controls

To ensure fair comparison, all models were evaluated using the same objectives, prompts, and artifacts within each control. Human assessors reviewed model outputs to confirm alignment with expected outcomes and to identify where professional judgment was still required to interpret evidence sufficiency, contextual factors, or partial compliance.

6. Observations

The evaluation results across the three RMF controls (PL-2, AU-6, and SR-6) revealed clear differences in performance among the six LLMs. Model outputs were assessed across accuracy of conclusions, gap detection, and reasoning quality, as defined by the evaluation criteria. In general, GPT-4, Claude, Gemini, and Copilot provided more consistent decision-support outputs, while performance varied by model in consistency and style of reasoning, with some models showing greater variability by control and artifact condition.

Overall Performance Patterns

Across all controls, models demonstrated baseline capability in summarizing control intent, restating requirements, and identifying obvious deficiencies in documentation. However, substantive differences emerged in how models translated observations into compliance determinations, weighted partial or implied evidence, and explained the rationale behind conclusions.

These differences became more pronounced as control complexity increased and as artifact quality degraded. Figure 1 summarizes average scores by model across accuracy, gap detection, and reasoning quality.

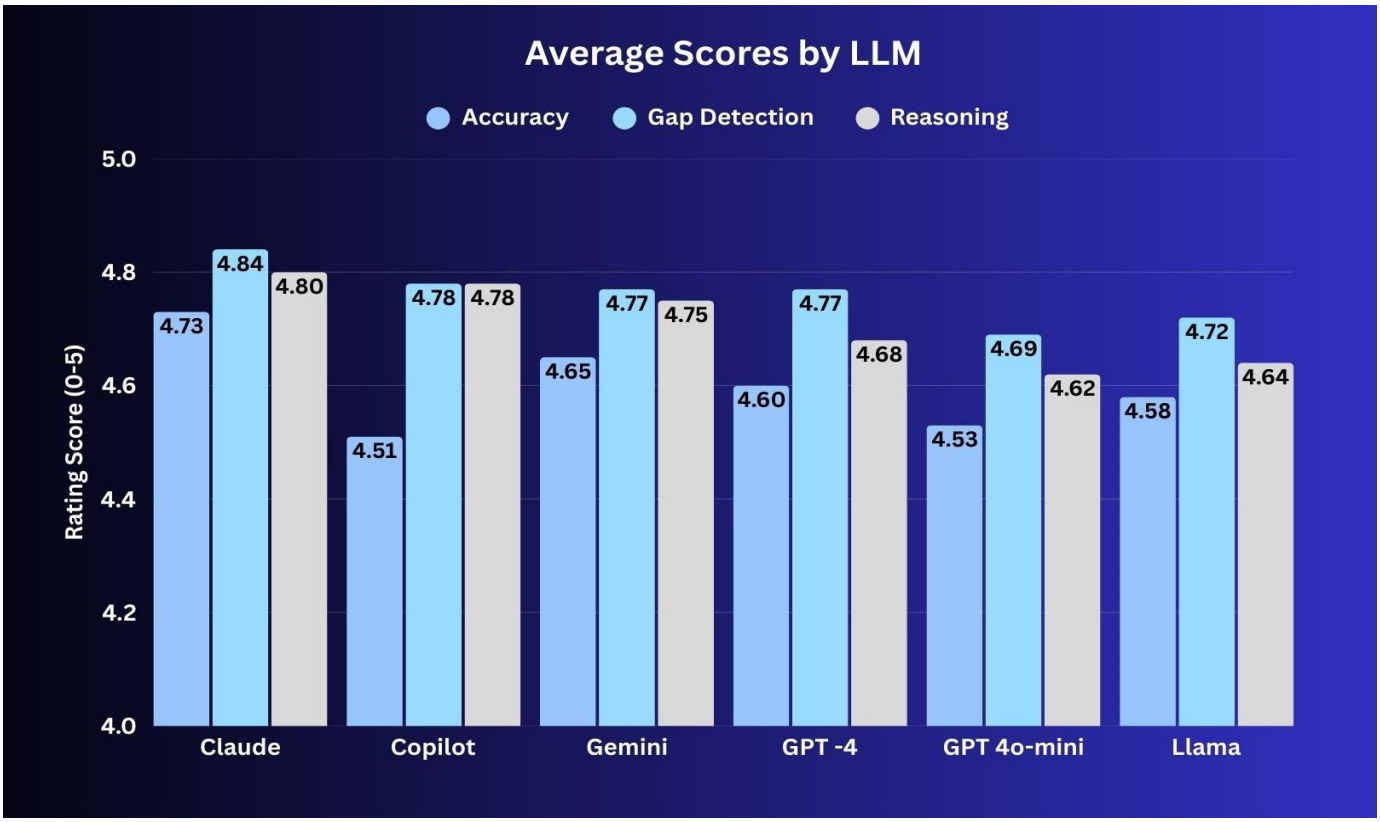

Figure 1 - Observed Average Scoring Across Model Performance.
This figure summarizes average model performance across all evaluated controls (PL-2, AU-6, SR-6), disaggregated by accuracy, gap detection, and reasoning quality.

Accuracy of Assessment Conclusions

Accuracy was measured based on alignment with expected human assessor determinations for the same artifact set.

Across controls:

Across controls, we observed greater score variance across models for AU-6 and PL-2, whereas SR-6 evaluations were more consistent. This suggests that models currently assess documentation-heavy controls like SR-6 more reliably than operationally or behavior-driven controls.

- AU-6 demonstrated the highest overall accuracy and consistency.
- SR-6 showed strong but tightly clustered performance across models.
- PL-2 exhibited moderate variability, particularly under narrative-heavy artifact conditions.

Average accuracy scores across models ranged from 4.71 (Claude) to 4.12 (GPT-4o-mini). While these differences were measurable, they remained within a compressed range, indicating that all models were generally capable of producing conclusions aligned with human expectations under structured prompts.

A key pattern also emerged in which models performed more consistently when artifacts contained explicit procedural structure and clear evidence of execution. Narrative-heavy documentation increased the likelihood of over-crediting

implementation strength. This suggests that model sensitivity to documentation style, not just content, can influence compliance determinations.

Gap Detection Capability

Gap detection scores were consistently high across all models, with averages ranging from 4.96 (Claude) to 4.76 (Llama).

Across all controls, models mostly:

- Identified missing or incomplete procedural elements
- Flagged vague or undefined roles and responsibilities
- Highlighted documentation inconsistencies
- Suggested areas requiring clarification

Gap detection exhibited less dispersion across models than accuracy or reasoning, indicating that identifying deficiencies may be a more stable capability across modern LLM architectures than translating those deficiencies into final compliance determinations.

Notably, improvements observed when moving from weak to strong artifacts were more pronounced than differences between individual models, indicating that artifact quality was a stronger driver of output quality than model selection.

Example:

The example below illustrates how the same model responds differently to identical assessment prompts when provided with strong versus weak evidence artifacts.

Prompt Question:

> 10. Do the procedures define a specific audit log review frequency (e.g., daily, weekly, monthly)? If yes, cite the relevant section.

Strong Artifact Response:

> 10. **Yes**, the procedures define a specific audit log review frequency (e.g., daily, weekly, monthly). According to Section 4.2, "ISAs conduct weekly reviews of audit dashboards and reports..."

Weak Artifact Response:

> 10. **No**, the procedures do not define a specific audit log review frequency (e.g., daily, weekly, monthly). Section 4 states that "Frequency of review depends on staff availability and system needs," which is vague and does not provide a clear schedule for log reviews.

This contrast highlights how artifact specificity and procedural clarity directly influence assessment outcomes, independent of the underlying model used.

Reasoning Quality

Reasoning quality produced the most meaningful differentiation across models.

Average reasoning scores ranged from 4.98 (Claude) to 4.73 (GPT-4o-mini). Although still tightly clustered, reasoning variation was more pronounced than accuracy variation.

Higher-scoring models demonstrated:

- Clear linkage between observed evidence and conclusions
- Explicit articulation of assumptions
- Calibrated confidence under weak artifact conditions
- Conditional phrasing when evidence was incomplete

Lower-scoring outputs more frequently:

- Collapsed partial implementation into binary determinations
- Presented conclusions without explicit evidentiary justification
- Failed to distinguish between documentation sufficiency and operational enforcement

These findings suggest that explanation depth and analytical transparency may be a stronger discriminator of model utility in RMF contexts than raw accuracy alignment alone.

Control-Specific Patterns

AU-6 (Audit Review, Analysis, and Reporting)

AU-6 required evaluation of audit frequency, role clarity, tooling, and escalation mechanisms. Models consistently identified gaps related to undefined review intervals, incomplete escalation paths, and ambiguous ownership. This control highlighted strengths in detecting procedural vagueness, along with a tendency in some outputs to infer operational maturity from partial detail. Score dispersion was slightly wider in AU-6, suggesting that procedural interpretation tasks may reveal clearer differentiation in reasoning styles.

PL-2 (System Security and Privacy Plans)

PL-2 emphasized documentation completeness and alignment with stated requirements. Models performed well in evaluating structural alignment but occasionally treated well-written narrative documentation as stronger evidence of implementation maturity than the artifacts supported.

SR-6 (Supplier Assessments and Reviews):

SR-6 required contextual evaluation of supplier lifecycle practices and supply chain risk management. While models successfully identified missing assessment steps or incomplete supplier monitoring practices, they varied in how they weighed lifecycle dependencies and third-party oversight nuance. Interestingly, SR-6 exhibited the tightest score clustering across models, suggesting that when evaluation criteria are structurally clear, model outputs may converge more closely.

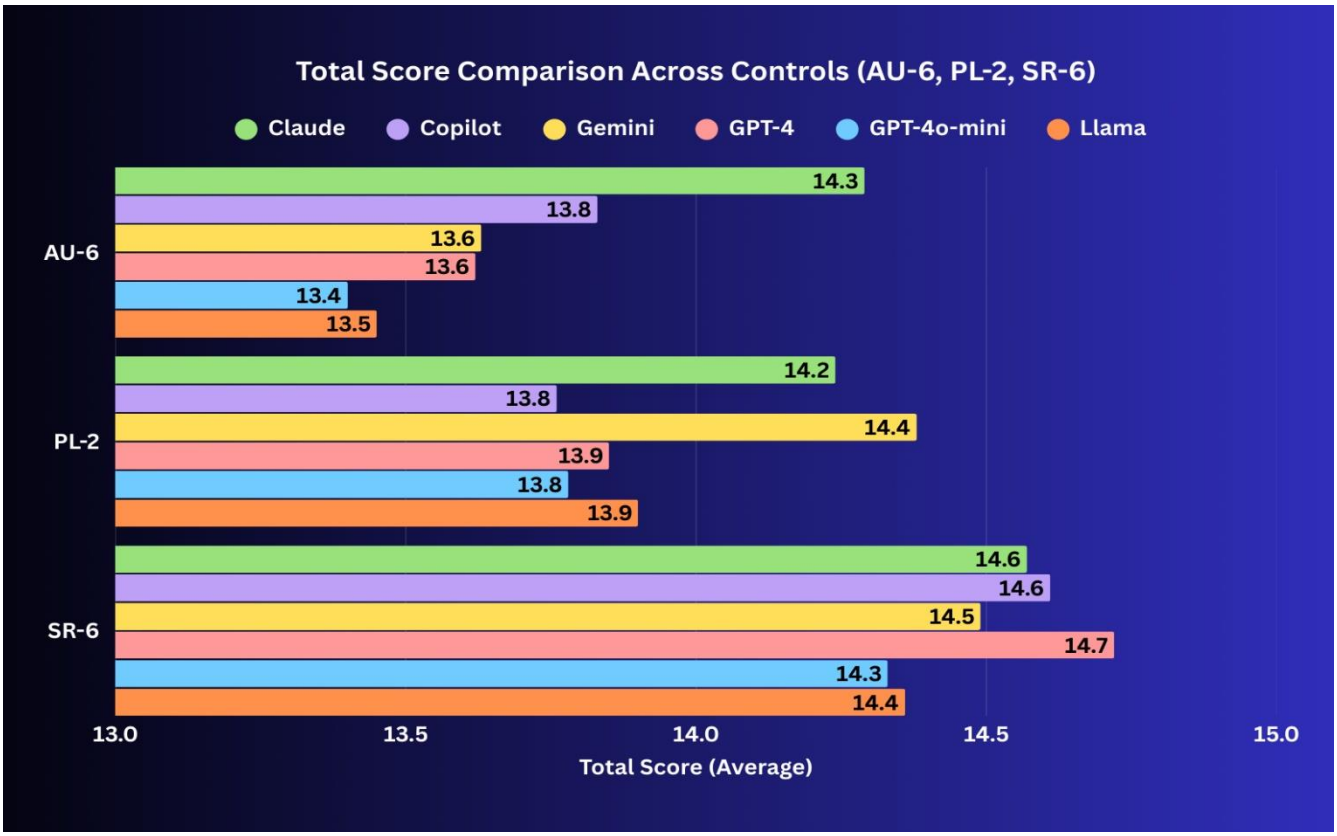

Figure 2 - Comparative performance patterns across PL-2, AU-6, and SR-6.
This figure illustrates the total scores of each model across the different controls (AU-6, PL-2, and SR-6).

Gap Identification vs. Compliance Determination

Across all models, a consistent distinction emerged between gap identification and compliance determination.

Models were generally effective at:

- Identifying missing elements
- Highlighting vague language
- Suggesting areas for improvement

Models were less consistent when:

- Translating gaps into pass/fail determinations
- Distinguishing between partial and insufficient implementation
- Applying assessor judgment where evidence was mixed

This demonstrates the role of AI as a decision-support tool, rather than a replacement for assessor judgment, particularly where control effectiveness depends on contextual interpretation.

Consistency and Variability Across Models

Differences between models were more apparent in:

- Depth and structure of reasoning
- Clarity of explanation
- Confidence expressed in conclusions

Despite these differences, overall assessment trends were largely consistent across models within the same control and artifact condition, suggesting that evaluation outcomes were driven more by task design and evidence quality than by model-specific behavior alone.

7. Key Takeaways

This evaluation highlights several practical considerations for agencies exploring the use of AI to support RMF assessment activities.

- AI models can meaningfully support assessors in summarizing documentation and identifying obvious gaps, particularly when evidence is well-structured.
- Evidence quality is the most significant driver of assessment outcomes. Well-developed artifacts had a greater impact on performance than differences between individual models.
- Models were generally more reliable in identifying deficiencies than in translating those deficiencies into compliance determinations.
- Human assessor judgment remains essential, particularly when evidence is incomplete, mixed, or context dependent.
- Differences between models were generally less significant than differences in task design, prompt structure, and artifact quality.
- Structured evaluation frameworks are key to preventing overreliance on AI-generated outputs.
- AI is best used as a decision-support capability within RMF programs, not as a replacement for professional judgment.

8. Limitations and Considerations

This evaluation was designed to provide practical insight into AI-supported RMF assessment activities. However, several factors should be considered when interpreting the findings.

- The evaluation focused on a limited set of controls and does not represent the full scope of NIST SP 800-53 requirements.
- Testing was conducted using simulated and de-identified artifacts, which may not capture all complexities of operational environments.
- Model performance was assessed in controlled testing scenarios rather than within live assessment workflows.
- Results reflect a point-in-time evaluation and may change as AI models and supporting platforms evolve.
- Human assessor judgment was used as a reference point, introducing inherent variability in expected outcomes.

9. Conclusion

This evaluation demonstrates that AI language models can provide meaningful support for RMF assessment activities when applied within structured, well-governed processes. When paired with high-quality evidence and informed human oversight, these tools can enhance consistency, efficiency, and analytical depth across assessment workflows.

The findings reinforce that AI systems are not substitutes for professional judgment. Variability in reasoning, sensitivity to evidence quality, and limitations in contextual interpretation underscore the continued importance of experienced assessors in authorization and continuous monitoring decisions.

As agencies continue to explore AI-enabled cybersecurity capabilities, disciplined evaluation frameworks and transparent governance practices will be essential to responsible adoption. Ongoing experimentation, practitioner engagement, and iterative refinement will be critical to realizing the benefits of AI while maintaining trust and accountability. Building on this evaluation, NR Labs is extending this work through the development of standardized prompt packs and agent-assisted assessment workflows, enabling more end-to-end evaluation of artifacts, structured findings, and draft POA&M entries. Ongoing work is examining the impact of model-specific prompt optimization to assess performance under maximized conditions.

Through its Cyber Innovation Practice, NR Labs will continue to evaluate emerging AI capabilities in partnership with practitioners and stakeholders to support evidence-based, mission-aligned cybersecurity programs.