SYM-AM-25-316

Excerpt from the Proceedings

of the

Twenty-Second Annual Acquisition Research Symposium and Innovation Summit

Wednesday, May 7, 2025 Sessions

Volume I

Leveraging Generative AI for Validating the Quality of DoD Acquisition Packages and Contract Documents

Published: May 5, 2025

Disclaimer: The views represented in this report are those of the author and do not reflect the official policy position of the Navy, the Department of Defense, or the federal government.

Approved for public release; distribution is unlimited.

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School

Leveraging Generative AI for Validating the Quality of DoD Acquisition Packages and Contract Documents

Samantha Nangia—is a Senior Program Analyst and the Enterprise Procurement System (ePS) Solution Manager for the Department of the Navy, overseeing the transformation of all Navy and Marine Corps procurement capabilities. She leads efforts to enhance procurement data quality, visibility, and integration, supporting enterprise analytics and auditability objectives. As the Lead Product Owner for the Navy's Electronic Procurement System (ePS), she ensures the delivery of key capabilities to the 1102 community. Nangia specializes in Procure-to-Pay (P2P) systems, data architecture, and automation, driving the integration of procurement and financial data. She has successfully automated the Procurement Performance Management Assessment Program (PPMAP) and developed solutions to improve procurement data quality across the Department of the Navy. She holds a bachelor's in computer science from Rollins College and an MBA from Georgetown University. Nangia's expertise in procurement systems and data solutions has contributed to improved efficiency and performance across the DON.

Tom Wardwell—is the Deputy Director of eBusiness Policy and Oversight serving as the Acquisition IT Portfolio Manager and eBusiness Transformation lead for the Department of the Navy. In this role, he oversees the strategic integration and management of a $200 million/year IT acquisition business systems portfolio, ensuring sound investment in capabilities supporting the acquisition mission and alignment with Navy-wide strategic objectives. He is critical in driving innovation, enhancing procurement efficiency, and ensuring compliance with evolving regulatory frameworks. He holds a bachelor's in physical science from the U.S. Naval Academy and master's degrees in acquisition management from the Naval Postgraduate School, information systems technology from The George Washington University, and national strategic resourcing from the Industrial College of the Armed Forces.

Randall Mora—is President and CEO of Avum, Inc., providing Avum's strategic and management direction and actively participates in multiple U.S. Navy, U.S. Army, U.S. Air Force, and the Fourth Estate (Department of Defense) Procure-To-Pay initiatives. His passion is leading Avum's research and development efforts to rapidly build solutions that Avum's customers can leverage for their success.

Carlos Parada Jr.—is a data scientist and statistician working with Avum, Inc. to provide high-quality products using the latest insights in machine learning and AI. Parada completed his degree in mathematics at Carleton College. He started in Bayesian inference and econometrics as a research developer with the Cambridge Machine Learning Group on the Turing.jl project. Since then, Parada has worked on medical trials testing cancer drugs and mass healthcare interventions in Cameroon. Parada's free time is often occupied by exploring emerging research within machine learning, artificial intelligence, and mathematics, coupled with his work on Wikipedia's mathematics articles.

Abstract

The Department of Defense (DoD) contracting process requires rigorous validation to ensure regulatory compliance, accuracy, and completeness. This paper explores the integration of NIPR GPT, a secure generative artificial intelligence (AI) model, to enhance the efficiency and reliability of the Acquisition and Contracting package validation. Deployed in a DoD-approved environment, NIPR GPT is a Government R&D Platform for GenAI models and applications serving as a comprehensive AI research and development platform featuring retrieval augmented generation. NIPRGPT enables model evaluation, shared workspaces, and secure document processing workflows, in our use case, we used it to automate key tasks such as compliance checks against FAR/DFARS/NMCARS/Local Policy Language, clause verification, contract risk identification, and data consistency validation. The proposed framework enables contracting officers to upload documents, select validation tasks, and receive detailed, actionable reports. NIPR GPT is able to leverage fine-tuned training on DoD-specific datasets to identify missing clauses, resolve ambiguities, and flag high-risk elements. By automating labor-intensive tasks, the system is able to reduce human error, accelerate processing, and ensure compliance with regulatory and policy requirements. The model is implemented within an IL-4 environment to address security concerns, with robust encryption protocols and access controls to safeguard sensitive data. Audit logging provides transparency, ensuring outputs can be reviewed and verified. A case study using a significant Aircraft procurement demonstrates the practical application of this framework. NIPR GPT identified missing compliance language and clauses, flagged ambiguous deliverable descriptions, and recommended corrective actions, streamlining the package approval process. This integration of AI into DoD workflows illustrates its potential to modernize procurement practices, improve accuracy, and maintain compliance in a highly regulated environment. This abstract highlights the transformative role of generative AI in supporting DoD contracting officers by providing reliable, secure, and efficient tools for package validation.

Keywords: Defense acquisition, Generative AI, LLM, NIPR GPT, FAR compliance, DFARS compliance, Contract validation, NMCARS, Procurement Performance Management Assessment Program, PPMAP, Regulatory automation, Compliance assurance, AI accuracy, Clause verification, Contracting workflows, FAR/DFARS interpretation, Human-in-the-loop AI, Contract analysis, Procurement automation, DoD contracting packages, DOD acquisition packages, Compliance errors, Automated legal review, Fine-tuned AI models, Regulatory compliance systems, AI benchmarking, Retrieval-augmented generation, RAG

Introduction

Background

The Department of Defense (DoD) contracting process is governed by a complex framework of regulations, including the Federal Acquisition Regulation (FAR), the Defense Federal Acquisition Regulation Supplement (DFARS), and the Navy and Marine Corps Acquisition Regulation Supplement (NMCARS). These regulations ensure that all contracts meet strict standards for compliance, accuracy, and completeness. For example, under DFARS 252.204-7012, contractors must implement cybersecurity measures to protect controlled unclassified information. Similarly, FAR Part 15 outlines detailed procedures for contract negotiations, ensuring fairness and transparency in source selection. NMCARS supplements these regulations by providing specific guidance for Navy and Marine Corps acquisitions, such as stricter validation of cost estimates and contract requirements. Rigorous validation processes, including proposal audits, compliance reviews, and independent cost estimates, help mitigate risks and ensure compliance with regulatory requirements while maintaining procurement integrity.

Problem Statement

Due to the complexity and evolving nature of federal acquisition regulations, contracting officers must navigate an intricate compliance landscape that includes the FAR, the DFARS, the NMCARS, and Department of the Navy (DON) policy directives. Each of these frameworks imposes stringent requirements on procurement processes, ranging from cost estimation and cybersecurity compliance to small business set-asides and contract auditing. The frequent updates and nuanced interpretations of these regulations add another layer of difficulty, increasing the risk of non-compliance, bid protests, and potential contract delays.

In response to these challenges, emerging generative AI models like NIPR GPT promise to automate regulatory analysis, reduce administrative burdens, and improve contract review efficiency. However, despite their potential, these models have not undergone rigorous validation to ensure their accuracy, reliability, and ability to identify critical compliance risks effectively. Errors in AI-generated recommendations could lead to overlooked compliance issues, misinterpretations of regulatory language, or unintended contract violations, ultimately jeopardizing mission readiness and procurement integrity. As such, there is a critical need to assess the efficacy of AI-driven solutions in real-world DoD contracting environments to determine their feasibility, limitations, and potential role in enhancing regulatory compliance.

Research Question/Objective

1. How accurately can NIPR GPT detect missing or misapplied language in DoD acquisition/contracting packages compared to traditional manual review processes?
2. What are AI-generated outputs' most common errors and limitations when validating regulatory and policy compliance?
3. To what extent does domain-specific fine-tuning improve the accuracy and reliability of NIPR GPT in assessing compliance with FAR, DFARS, NMCARS, and DON policy directives?
4. How can retrieval-augmented generation (RAG) improve the accuracy and completeness of AI-driven compliance validation in DoD acquisition/contracting packages?

This research aims to evaluate the effectiveness of AI-driven solutions, particularly NIPR GPT and RAG, in enhancing compliance validation for DoD contracting. The study focuses on assessing accuracy, identifying limitations, measuring the impact of fine-tuning, and exploring the potential benefits of integrating AI into existing procurement workflows. The key objectives are as follows:

1. **Evaluate AI Accuracy**: Assess NIPR GPT's effectiveness in identifying compliance issues compared to manual review.
2. **Identify AI Limitations**: Analyze common errors and gaps in AI-generated compliance assessments.
3. **Measure Fine-Tuning Impact**: Determine how domain-specific fine-tuning improves AI performance in regulatory compliance tasks.
4. **Compare Review Efficiency**: Investigate whether AI-assisted reviews can reduce the time and effort required for compliance validation while maintaining accuracy.
5. **Assess RAG Effectiveness**: Evaluate how RAG enhances AI-generated outputs by integrating real-time regulatory references.
6. **Analyze Error Reduction**: Examine whether RAG reduces common AI errors, such as hallucinations, misinterpretations, or outdated regulatory references.
7. **Optimize AI Integration**: Identify best practices for implementing AI-driven compliance validation in DoD procurement workflows.

Scope and Limitations

Scope: This research examines the potential of AI-driven compliance validation in DoD procurement, focusing on NIPR GPT and RAG to enhance accuracy, efficiency, and regulatory adherence. The study aims to:

1. Assess the accuracy of AI in detecting missing or misapplied regulatory language compared to traditional manual review processes.
2. Identify common errors and limitations in AI-generated compliance assessments.
3. Evaluate the impact of domain-specific fine-tuning on the AI model's ability to interpret and apply FAR, DFARS, NMCARS, and DON policy directives.
4. Investigate how RAG improves the accuracy and completeness of AI-generated outputs by integrating external regulatory sources.

5. Explore the feasibility of implementing AI-assisted compliance validation in real-world DoD procurement workflows.

The research will utilize a dataset of DoD acquisition packages and contract documents to test AI performance. It will also include qualitative insights from Acquisition professionals to understand AI's practical applications and limitations.

Limitations: While this study provides valuable insights into AI-driven compliance validation, certain limitations exist:

1. *Data Availability*: This research relies on contract data accessible within the DON enclaves, which are not publicly available. While this controlled environment ensures data security and regulatory compliance, it may limit the diversity of acquisition packages used to evaluate AI performance. As a result, findings may not fully account for the variability in contract structures across different DoD agencies or broader procurement scenarios.
2. *Regulatory Updates*: AI models may not instantly adapt to evolving regulatory changes, impacting the accuracy of compliance validation over time.
3. *AI Interpretability*: NIPR GPT's decision-making process may lack transparency, making it challenging to fully understand how compliance determinations are made.
4. *Scope of Fine-Tuning*: The study focuses on domain-specific fine-tuning but does not explore real-time learning or continuous retraining of AI models.
5. *Human Oversight*: AI is not intended to replace human acquisition professionals but to assist them. The research does not propose a fully automated compliance validation system but rather a hybrid approach where AI enhances manual reviews.
6. *Comparative Baseline*: Manual review processes vary across contracting offices, which may introduce inconsistencies when comparing AI performance against human evaluations.

By defining these scope and limitations, this research ensures a focused and realistic assessment of AI's potential in DoD contracting compliance validation.

Methodology

This research employs a multi-faceted methodology to evaluate AI-driven compliance validation in DoD contracting. By leveraging structured compliance checks, RAG, and expert feedback, we aim to assess the accuracy, reliability, and practical applicability of AI models in regulatory reviews. The key components of our methodology include:

1. **Structural Compliance Checks:**

    - *Annex 18 ISTRAPS, Annex 19 PSTRAP-M, and Annex 20 ISTRAP-M:* Evaluate structural compliance of acquisition packages to ensure adherence to regulatory frameworks.

    - *Annex 1 Review of Justifications & Approvals (J&As)*: Assesses the logic, strength of argument, and flow of J&As. AI-generated prompts guide reviewers in refining responses, ultimately producing revised J&As.

2. **Contract Award Document Analysis:**
    - Compares final contract award documents against FAR, DFARS, and NMCARS to identify structural compliance gaps, risks, clause compliance analysis, and regulatory adherence.
    - Utilizes PDS XML data to standardize and enhance acquisition package analysis.
3. **RAG Model for Compliance Checks:**
    - Uses “clean” acquisition packages—deemed high-quality by the team—as reference data to improve AI-generated compliance recommendations.
    - Assess the impact of RAG-enhanced AI outputs on the accuracy and completeness of contract reviews.
4. **Expert Review and Model Refinement:**
    - Incorporates DASN (P) and legal review comments to fine-tune AI-generated outputs through prompt engineering.
    - Evaluates the fidelity of AI-generated compliance assessments against expert feedback.
5. **Batch Summarization and Ranking:**
    - Processes multiple acquisition packages simultaneously to generate structured summaries.
    - Ranks packages based on quality and provides recommendations for strengthening weaker submissions.

This methodology ensures a rigorous and iterative approach to assessing AI’s potential in DoD contracting, balancing automation with expert validation to enhance regulatory compliance and efficiency.

NMCARS Analysis

To assess structural compliance in DoD acquisition, we collaborated with the Program Analytics Business Transformation (PABT) team at DASN (P) to obtain contract review samples and results. This dataset includes both annotated assessments and high-quality “clean” contract versions, providing a foundation for AI model evaluation. The specific annexes reviewed include:

- Annex 1—Review of J&As: Assesses logic, argument strength, and flow; AI-generated prompts guide reviewers in refining responses, leading to revised J&As.
- Annex 18—Individual Streamlined Acquisition Plan (ISTRAP): Evaluates structural compliance in acquisition planning.
- Annex 19—Program Streamlined Acquisition Plan With Services (PSTRAP-M): Focuses on acquisition planning for service-based contracts.
- Annex 20—Individual Streamlined Acquisition Plan With Services (ISTRAP-M): Examines structural compliance for individual acquisition plans involving services.

These structured compliance checks provide a benchmark for AI-driven analysis, enabling comparison between AI-generated outputs and expert-validated contract reviews. The results inform subsequent model fine-tuning and RAG-based improvements for enhanced regulatory compliance validation.

This ensures clarity and emphasizes how the collaboration with the PABT team strengthens the research methodology.

RAG-Based Improvements for Enhanced Compliance Analysis

We leveraged RAG within the NIPR GPT large language model (LLM) to improve the accuracy and completeness of AI-driven compliance validation (Google Cloud, n.d.). By incorporating high-quality contract documents and structured compliance assessments into the retrieval system, we aim to enhance AI-generated compliance recommendations.

This process involves:

- **Feeding validated contract reviews and "clean" documents** into the RAG model as reference materials.
- **Enhancing AI outputs** by dynamically retrieving relevant regulatory information from FAR, DFARS, NMCARS, and DON policy directives.
- **Comparing AI-generated compliance assessments** before and after RAG integration to measure accuracy, completeness, and risk identification improvements.

As illustrated in Figure 1, RAG-based retrieval enables the AI model to access real-time reference data, reducing errors such as misinterpretations and hallucinations while improving the overall quality of compliance validation (Zvornicanin, 2024).

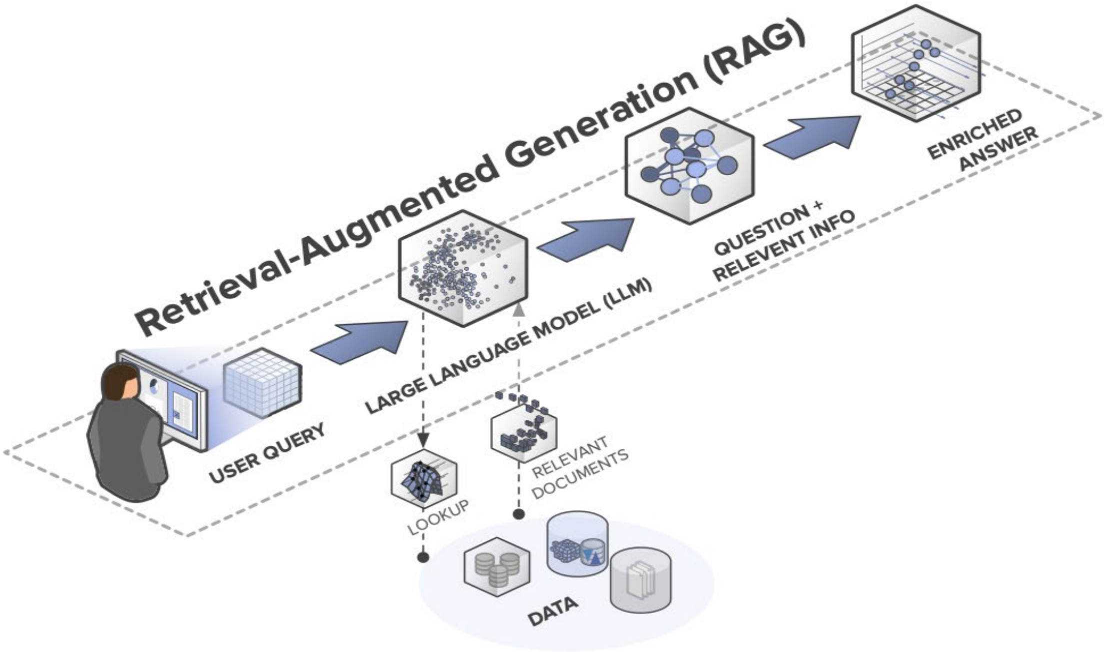


Figure 1. RAG-Enhanced Compliance Validation Workflow

FAR/DFARS/NMCARS Clause Compliance Analysis

This research utilizes the Procurement Data Standard (PDS) XML as the primary data source to systematically assess regulatory compliance, risk factors, and structural integrity in

DoD contracting (OUSD[A&S], n.d.-a). PDS is a standardized XML format that captures structured procurement information (OUSD[A&S], n.d.-b), enabling automated validation against **FAR, DFARS, NMCARS**, and DON policy directives. By leveraging PDS, we aim to:

- Standardize contract data ingestion for AI-driven compliance analysis.
- Identify structural inconsistencies, regulatory gaps, and missing clauses in contract award documents.
- Develop intelligence models that enhance AI's ability to detect compliance risks, misapplied, and missing clauses.
- Improve automation and accuracy in contract validation by integrating structured PDS data into AI and RAG models (Atamel, 2025).

Our goal is to ensure a consistent, scalable, and **data-driven approach** to evaluating compliance across diverse procurement scenarios by aligning contract award document analysis with PDS XML.

While PDS XML provides a standardized structure for analyzing contract award data, it has inherent limitations when used for comprehensive clause compliance validation. The XML primarily reflects explicitly listed clauses and those tagged for inclusion, but it does not account for clauses incorporated by reference in attachments or supplemental documents.

As such, during our analysis, clause presence or absence is inferred based on what is available in the XML structure. The system may flag potential omissions, but it is important to note:

> "Based on the provided XML and a general understanding of government contracting, certain clauses may appear to be missing. However, we cannot definitively determine omission without access to the full contract file, including all attachments and incorporated references. Some clauses may exist elsewhere in the contract documentation but are not surfaced in the PDS XML."

To mitigate this limitation, our methodology incorporates:

- Subject matter expert (SME) validation of flagged discrepancies,
- Pattern recognition from previously reviewed complete packages,
- And continuous refinement of the model prompts to account for common clause placement practices.

This ensures that while the system offers valuable insights into potential compliance issues, final validation still benefits from expert oversight and full contract context.

Key components of this approach include:

- **Comparative Clause Analysis**—Evaluating Clause Logic Service (CLS) recommendations against actual contract clauses to identify discrepancies, misapplications, and gaps.
- **Benchmarking "Clean" Contracts**—Developing a validated reference set of high-quality contract awards, serving as a compliance standard for future assessments.
- **AI-Enhanced Compliance Checks**—Integrating RAG within NIPR GPT to:

    - Analyze new contract awards against the clean contract benchmark.
    - Improve clause validation accuracy by refining AI-driven assessments (Zvornicanin, 2024).
- **Expert Review and Model Refinement**—Leveraging the expertise of experienced contracting officers, DASN (P) analysts, and legal professionals to:
    - Incorporate DASN (P) and legal review comments to fine-tune AI-generated outputs through prompt engineering.
    - Evaluate the fidelity of AI-generated compliance assessments against expert feedback to improve model accuracy and reliability (Atamel, 2025).

By combining automation with expert judgment, this methodology ensures efficient, accurate compliance validation while reinforcing the essential role of contracting officers and subject matter experts. The goal is to streamline regulatory adherence in DON contracting by reducing manual review efforts, improving compliance precision, and continuously refining AI-based assessments through SME-driven oversight.

Batch Summarization and Ranking

To enhance structural analysis and compliance validation across FAR, DFARS, and NMCARS, this research develops a Python-based process for batch summarization, ranking, and accuracy assessment of contract documents. A critical aspect of this methodology is detecting AI hallucinations, which occur when the model generates incorrect or misleading information. These errors can be identified using entropy, a measure of the model’s uncertainty in its predictions. High-entropy responses indicate that the model is uncertain about the correct answer, signaling a higher likelihood of inaccuracies or hallucinations (Entropy [information theory], n.d.).

The methodology consists of the following key components:

1. **Structural Analysis at Scale**—Processing multiple contract documents to assess completeness, clause structure, and alignment with regulatory requirements.
2. **Entropy-Based Accuracy Evaluation**—Assigning entropy scores to AI-generated outputs, where:
    - Low entropy suggests a high-confidence prediction.
    - High entropy signals uncertainty and a potential AI hallucination.
3. **Flagging and Visualizing High-Entropy Responses**—High-entropy answers are flagged for additional review and highlighted in the interface, enabling SMEs to focus on ambiguous or unreliable AI outputs.
4. **Ranking and Prioritization**—Summarized contract analyses are ranked based on compliance confidence, directing attention to high-risk discrepancies for further evaluation.

To substantially improve AI model performance, we integrate advanced reasoning techniques such as:

- Prompt engineering—Combining human-generated instructions with AI enhancements to provide additional clarity and encourage the model to “think things through” (OpenAI, 2024).
- Chain-of-Thought (CoT; Gadesha & Kavlakoglu, 2024)⎯Requiring the model to provide structured, step-by-step logical reasoning (Founding Minds, 2024; Villani, 2024).
- Beam search—Replacing the traditional “greedy sampling” with a more-advanced beam search optimization algorithm (Leblond et al., 2021).
- Structured generation (OpenAI, 2024)⎯Enforcing a prespecified output formatting (given by a context-free grammar such as JSON) to provide a “scaffolding” and restrict the AI to generating valid outputs.

By combining structured automation (Shorten et al., 2024), entropy-based validation (Entropy [information theory, n.d.), and expert oversight, this approach enhances compliance accuracy, reduces AI hallucinations, and improves regulatory adherence.

Results and Analysis

The results of our acquisition document/ package structural review stem from the integration of annotated acquisition documents provided by the DASN (P) PABT team with our RAG model analysis in NIPR GPT. These annotated documents, sourced from previous acquisition package reviews, contained NMCARS sections mapped directly to contract content, serving as a ground truth dataset for evaluating and improving structural compliance validation.

The annotated acquisition documents were ingested into our RAG pipeline, enabling the model to retrieve contextually relevant regulatory references when analyzing contract clauses.

Key steps and observations are included below.

Basic Use Case—Annex 18—ISTRAP Structural Review

Doc 1: TAB A - NMCARS 18-25 Annex 18 - ISTRAP. This is the template for an ISTRAP and the rules to be followed. In this document, we provided the model for the NMCARS structure to analyze the remaining against.

Doc 2: ISTRAP CNO Avails Fast Attacks Sub. This is the first document we assess against the template for structural/content compliance. It contains three tabs—each tab is a different version of the same document—(B1) initial submission; (B2) DASN Edits; (B3) Final Clean Copy.

Doc 2-B1: TAB B1 - ISTRAP CNO Avails Fast Attacks Sub

Doc 2-B2: TAB B2 - ASN(RDA) CR_TC_ISTRAP CNO Avails

Doc 2-B3: TAB B3 - ISTRAP CNO Avails Fast Attacks Sub_20241213 CLEAN

Doc 3: DDG-FFG PY Acquisition Plan v9. This is the second document we use to assess against the template for structural/ content compliance. It contains three tabs—each tab is a different version of the same document—(C1) initial submission; (C2) DASN Edits; (C3) Final Clean Copy.

Doc 3-C1: TAB C1 - ISTRAP CNO Avails Fast Attacks Sub

Doc 3-C2: TAB C2 - ASN(RDA) CR_TC_ISTRAP CNO Avails

Doc 3-C3: TAB C3 - ISTRAP CNO Avails Fast Attacks Sub_20241213 CLEAN

Doc 4: DDG-FFG PY Acquisition Plan v9. This is the third document we use to assess against the template for structural/content compliance. It contains three tabs—each tab is a different version of the same document—(C1) initial submission; (C2) DASN Edits; (C3) Final Clean Copy.

Doc 4-D1: TAB D1 - ISTRAP CNO Avails Fast Attacks Sub

Doc 4-D2: TAB D2 - ASN(RDA) CR_TC_ISTRAP CNO Avails

Doc 4-D3: TAB D3 - ISTRAP CNO Avails Fast Attacks Sub_20241213 CLEAN

Summary Result of ISTRAP Structural Review Use-Case

Across three iterations using NIPR GPT, the model's assessment of the ISTRAP showed alignment with formal reviewer comments, though weaknesses and areas for refinement were identified. Its capacity for rapid initial assessments promises to enhance both the efficiency of the contract writing process and the accuracy of acquisition package compliance.

Enhanced Use Case—Annex 1—J&A Structural and Logical Flow Review

Doc 1: TAB A - NMCARS 18-25 Annex 1 - J&A. This is the template for a J&A and rules to be followed. We feed the model this document to assess the rest against.

Doc 2: J&A CJA No. CR-24219. This is the first document we assess against the template for structural/content compliance. It contains three tabs—each tab is a different version of the same document—(B1) initial submission; (B2) DASN Edits; (B3) Final Clean Copy.

Doc 2-B1: TAB B1 - CJA No. CR-24219 - Body - MS Word

Doc 2-B2: TAB B2 - CJA No. CR-24219 - Track Changes - DASN

Doc 2-B3: TAB B3 - CJA No. CR-24219_Final Clean 22JAN25

Doc 3: Draper CPS JA23-51. This is the second document we use to assess against the template for structural/content compliance. It contains three tabs—each tab is a different version of the same document—(C1) initial submission; (C2) DASN Edits; (C3) Final Clean Copy.

Doc 3-C1: TAB C1 - Draper CPS JA23-51 Final CLEAN

Doc 3-C2: TAB C2 - 20240806_ASN(RD&A) CR_TC_Draper CPS JA23

Doc 3-C3: TAB C3 - Draper CPS JA23-51 Final CLEAN

Doc 4: NAVSEA JA_DDG91. This is the second document we use to assess against the template for structural/content compliance. It contains three tabs—each tab is a different version of the same document—(C1) initial submission; (C2) DASN Edits; (C3) Final Clean Copy.

Doc 4-D1: TAB D1 - NAVSEA JA_DDG91_DMP2_to_NASSCO

Doc 4-D2: TAB D2 - ASN(RDA) CR_TC_NAVSEA_DDG91 DMP J&A - NA

Doc 4-D3: TAB D3 - Final USS PINCKNEY (DDG 91) DMP JA 42,9

Draper J&A (SPJA23-51):
Total Comments in Initial Review: 5
Comments with Direct Alignment: 4
Proportion of Alignment: 4/5 = 80%

Analysis: There was a strong degree of alignment for the Draper J&A, indicating that NIPR GPT was generally on the right track in identifying the key areas of concern.

USS PINCKNEY (DDG 91) Modernization J&A (J&A Number: 42,916):
Total Comments in Initial Review: 7
Comments with Direct Alignment: 6
Proportion of Alignment: 6/7 = 86%

Analysis: The alignment was even stronger for the USS PINCKNEY J&A, suggesting that NIPR GPT's understanding of the review team's priorities had improved.

Structural Review—J&A Enhanced Use Case

To assess the models' capabilities in both structurally reviewing Navy acquisition packages and assessing the strength of the document's arguments through logical flow assessments, testing was conducted across Annex 1 (J&As). NIPR GPT, across multiple iterations, showed increasing degrees of alignment with formal reviews including legal reviews, reinforcing its potential as a supplementary tool that can significantly enhance acquisition efficiency and compliance accuracy.

Structural Review—Multiple Use Cases

To assess the models' capabilities in structurally reviewing Navy acquisition packages, testing was conducted across Annex 1 (J&As), Annex 18 (ISTRAP), Annex 19 (PSTRAP-M), and Annex 20 (ISTRAP-M). NIPR GPT, across multiple iterations, showed varying degrees of alignment with formal reviews, highlighting the need for refinement but reinforcing its potential as a supplementary tool that can significantly enhance acquisition efficiency and compliance accuracy. The following is an analysis of the alignment between NIPR GPT's comments and the formal review team's comments for each of the three J&A packages, expressed as a proportion of comments that aligned. As an example, this was the Annex 1 Quantitative Result.

Overall Trend

The proportion of alignment increased over the three J&A packages, demonstrating that the review process became more closely aligned with the review team's perspective. This suggests that NIPR GPT effectively learned from the previous reviews and incorporated that knowledge into the subsequent assessments.

Key Takeaways

- **Effective Learning:** The increasing proportion of alignment indicates that NIPR GPT was able to effectively learn from the review team's feedback and incorporate their priorities into its own review process.
- **Areas for Improvement:** Even with the high levels of alignment, there were still some comments that NIPR GPT missed. This highlights the importance of continuous learning and refinement of the review process.
- **Value of Different Perspectives**: The combination of the model's general assessment and the review team's specific comments resulted in a more comprehensive and robust evaluation of the J&As.

Clause Compliance PDS-Based Analysis Approach

The clause compliance portion of this research began by analyzing the PDS XML contract award data to identify potentially missing clauses, which were evaluated based on known regulatory requirements and contract characteristics (e.g., contract type, dollar value, and acquisition strategy).

Recognizing the limitations of PDS data—which may not include clauses incorporated by reference or detailed in attachments—our initial approach focused on detecting likely omissions based on what was explicitly represented in the XML.

To facilitate structured analysis, potentially missing clauses were categorized by key areas of regulatory concern, including but not limited to:

Cost and Pricing:

- **FAR 52.216-7 Allowable Cost and Payment:** While mentioned within another clause, it should likely be a standalone clause, especially in a CPFF contract.
- **FAR 52.216-8 Fixed Fee:** Similar to the above, while referenced, it's best practice to include it directly.
- Clauses related to cost accounting standards (CAS), if applicable to the contractor.

Changes and Terminations:

- **FAR 52.243-1 Changes—Fixed-Price:** Or the appropriate Changes clause for a CPFF contract if modifications are anticipated beyond the issuance of task orders.
- **FAR 52.249-2 Termination for Convenience of the Government (Fixed-Price):** Or the appropriate termination clause for a CPFF contract.

Data Rights and Intellectual Property:

- Specific data rights clauses (e.g., **DFARS 252.227-7013 Rights in Technical Data—Noncommercial Items**) define ownership and usage of technical data. The XML mentions a “data rights strategy,” making these clauses highly likely to be needed but potentially located elsewhere in the full contract.
- Clauses related to patents and copyrights, if applicable.

Subcontracts:

- **FAR 52.244-2 Subcontracts (Cost-Reimbursement and Letter Contracts):** Or the appropriate subcontracts clause for a CPFF contract.
- **FAR 52.219-9 Small Business Subcontracting Plan**, if applicable, based on the dollar value and nature of the work. The PWS mentions small business contracting, suggesting this clause might be necessary.

Other Important Areas (clauses may be needed depending on specific circumstances):

- **Inspection of Services:** A clause defining acceptance criteria and inspection procedures.
- **Insurance:** Clauses requiring specific types and levels of insurance.
- **Disputes:** A clause outlining the dispute resolution process.
- **Equal Opportunity:** Clauses related to equal employment opportunity and affirmative action.
- **Labor Standards:** Clauses related to labor laws (e.g., Service Contract Act, if applicable).

Key Takeaway

This list is not exhaustive and serves as a starting point. **We then consulted the full contract, all incorporated documents, and any applicable regulations (FAR, DFARS, NMCARS) to determine the complete set of required clauses to supplement the initial**

results, consulting with contract specialists for a thorough review. The model identified many applicable clauses and terms; however, ***not all*** left erroneous references and missing language.

Contract Risk and Compliance Assessment Using PDS XML

Next, the team provided PDS XML files representing multiple DON contract awards as part of the analysis. These structured data files were ingested by NIPR GPT, which performed a multi-layered review of each contract, focusing on both structural risks and regulatory compliance.

The system produced a detailed breakdown consisting of the following core elements: an overview of the contract's key terms, type, scope, and funding structure derived from the PDS metadata.

1. **Summary of the Contract**
2. **Risks Associated with the Contract Structure**
    - Cost-Plus-Fixed-Fee (CPFF) Risk
    - Indefinite Delivery Indefinite Quantity (IDIQ) Risk
    - Incrementally Funded Contract Risk
    - Level of Effort (LOE) Risk
    - Personnel Risks
    - Proprietary Information Agreements (PIAs) and Technical Assistance Agreements (TAAs) Risk
    - Travel Costs (Cost No Fee)
3. **Compliance with FAR, DFARS, and NMCARS Stipulations**

Contract Risk and Compliance Assessment NIPR GPT Output (*summary only*):

Based on the structured **PDS XML data** provided for each contract, **NIPR GPT** generated a detailed compliance assessment across multiple regulatory frameworks. The output was structured into the following categories:

1. **FAR Compliance**
   – Evaluation of required FAR clauses based on contract type, dollar value, and other contextual metadata.

2. **DFARS Compliance**
   – Assessment of DoD-specific regulatory provisions and how well they were represented in the contract data.

3. **NMCARS Compliance**
   – Review of DON-specific clauses and guidance under the NMCARS.

4. **Potential Compliance Concerns and Areas for Further Review**
   – Identification of clauses or contract features that may warrant additional scrutiny, including possible omissions or inconsistencies.

5. **Recommendations**
   – Actionable suggestions provided for both the **government** (e.g., clause corrections, structural risks) and the **contractor** (e.g., documentation improvements or clarifications).

A disclaimer accompanied each assessment to clarify the scope and limitations of the AI-generated analysis:

> “This analysis is based solely on the provided XML file. A complete assessment would require a review of the entire contract document, including all attachments and incorporated clauses. I am not a legal professional, and this is not legal advice.”

Summary: Clause Set Completeness and Future Research Direction

During the course of our research, it became clear that **complete clause visibility—** including clauses incorporated by reference or detailed in attachments—is **essential for accurate contract compliance analysis**. Relying solely on the PDS XML representation proved insufficient for a comprehensive assessment, as it omits many contextually critical clauses.

As a result, we identified a **promising direction for further research**: leveraging the Clause Logic Service (CLS) to **associate its recommended clause set with the actual clauses present in the final award PDS**. This involves:

- Pulling both the **CLS recommendation outputs** and the **final contract award PDS** data via interface integration
- Creating **mappings between recommended and actual clauses**
- Analyzing the differences to identify **compliance gaps and best practices**

Through this process, we aim to **identify high-quality, “clean” contract examples**—those that demonstrate strong alignment between CLS guidance and final execution. These exemplar contracts can then be used to **fine-tune our RAG model**, supporting improved clause prediction and validation in future awards.

Programmatic Interface for Batch Summarization and Ranking

We developed a programmatic interface to interact with an LLM to efficiently conduct automated document analysis, requirement-specific grading, and confidence scoring. This interface enabled batch processing of contract documents, allowing for structured assessments across multiple dimensions, including:

1. **Automated Document Analysis**—Parsing and analyzing contract documents at scale to identify structural integrity, clause completeness, and regulatory compliance.
2. **Requirement-Specific Grading**—Evaluating each package against predefined compliance criteria, assessing adherence to FAR, DFARS, and NMCARS.
3. **Confidence Scoring**—Assigning an entropy-based confidence score to each assessment, flagging high-uncertainty outputs for further expert review.
4. **Requirement Grading**—Ranking packages based on compliance strength, highlighting areas that require revision or additional scrutiny.

The programmatic interface facilitated seamless interaction with the LLM, ensuring a repeatable and efficient workflow for large-scale contract evaluations. We utilized obfuscated data to test and refine the automated interface, allowing for code validation and efficiency testing while preserving data integrity and security.

This interface streamlined contract analysis workflows by integrating batch processing, ranking mechanisms, and AI-driven confidence scoring, enhancing review accuracy, efficiency, and scalability. Future iterations will focus on further optimizing response accuracy, refining ranking methodologies, and enhancing real-time feedback mechanisms for contract evaluators. Key steps and observations from the development team include:

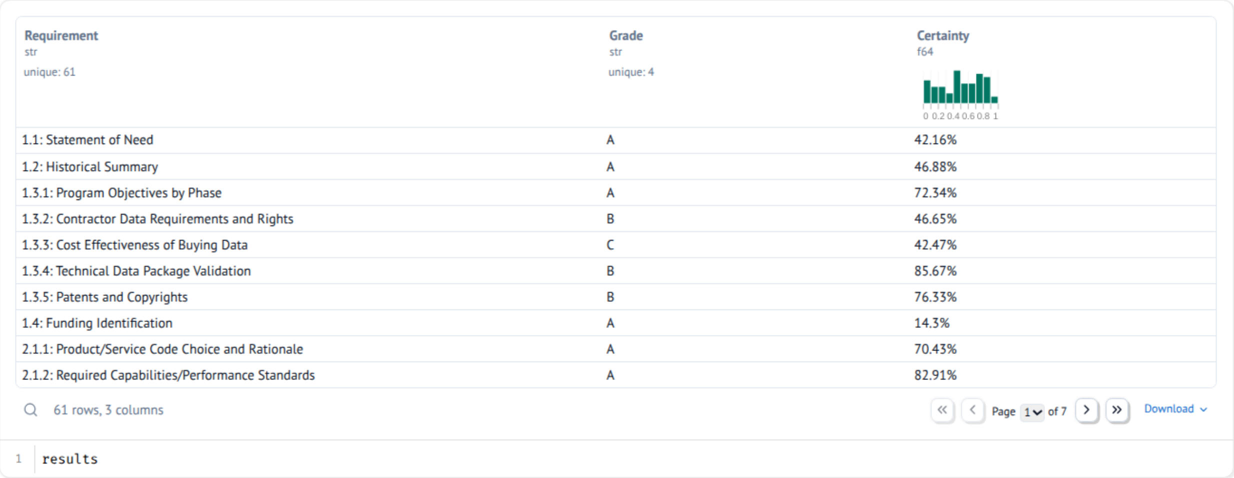
This framework uses ISTRAP documents against a predefined set of requirements to automate the analysis. Using the Gemini 2.0 Flash Lite AI model, each document is evaluated and graded (A–F) based on how well it meets each individual requirement. A certainty score, derived from the model's perplexity (exponentiated entropy), indicates the confidence level of each assigned grade (Tornetta, 2021):

$$C(X) := 1 - PP(x) = exp \sum_{i \in A:F} p_i \, ln \, p_i$$

The final output is a table showing the grade and certainty for each requirement in the analyzed document.

1. **Automated Document Analysis:** The software takes a set of ISTRAP documents (PDFs found in `data/contracts/`) and systematically analyzes each one against a predefined list of requirements derived from the `reqs/ISTRAP.pdf` and structured in `reqs/response-schema.txt`.
2. **Requirement-Specific Grading:** For *each* document, the AI model (Gemini 2.0 Flash Lite) is prompted to evaluate how well that document addresses *each individual requirement* listed in the schema (Liu et al., 2021).
3. **Grading Scale:** The AI assigns a grade (A, B, C, D, or F) for every requirement within each document, based on the instructions provided (`system_instruction`) and the specific requirement's description. Table 1, Grading Scale for ISTRAP Reviews, shows prioritized results for ISTRAP reviews, including model certainties in outcomes (measured using model perplexity over tokens).
4. **Confidence Score (Certainty):** The software calculates a "Certainty" score for each grade assigned from the model's perplexity (Morgan, 2024). Perplexity is a measure of how surprised or uncertain the model was when generating the response (the grade). Lower perplexity (closer to 1) means higher certainty/confidence in the assigned grade. Figure 2, Histogram Count of ISTRAP Requirements Colored by Grade, illustrates the model's confidence score for each ISTRAP requirement.
5. **Tabular Output (`results` DataFrame):** The final output is a table (currently showing results for *one* processed document). This table has three columns:
    - `Requirement`: The specific ISTRAP requirement text being evaluated (e.g., "1.1: Statement of Need").
    - `Grade`: The A–F grade assigned by the AI for that requirement in the analyzed document.
    - `Certainty`: The calculated confidence score for that specific grade.
6. **Overall Confidence Assessment:** An overall perplexity score is calculated across all the grades for the analyzed document, providing a general sense of the AI's confidence in its assessment of that entire document.

Table 1. Grading Scale for ISTRAP Reviews



Requirement str unique: 61	Grade str unique: 4	Certainty f64
1.1: Statement of Need	A	42.16%
1.2: Historical Summary	A	46.88%
1.3.1: Program Objectives by Phase	A	72.34%
1.3.2: Contractor Data Requirements and Rights	B	46.65%
1.3.3: Cost Effectiveness of Buying Data	C	42.47%
1.3.4: Technical Data Package Validation	B	85.67%
1.3.5: Patents and Copyrights	B	76.33%
1.4: Funding Identification	A	14.3%
2.1.1: Product/Service Code Choice and Rationale	A	70.43%
2.1.2: Required Capabilities/Performance Standards	A	82.91%

61 rows, 3 columns

```
results
```

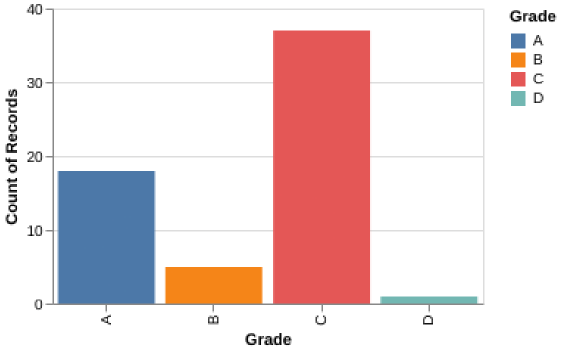


Figure 2. Histogram Count of ISTRAP Requirements Colored by Grade

Figure 3, Histogram Count of the Model Certainty Colored by Grade, illustrates ISTRAP graded point-by-point for compliance; we generate summary charts to help users understand both the evaluated quality of the ISTRAP and the model's confidence in its outputs.

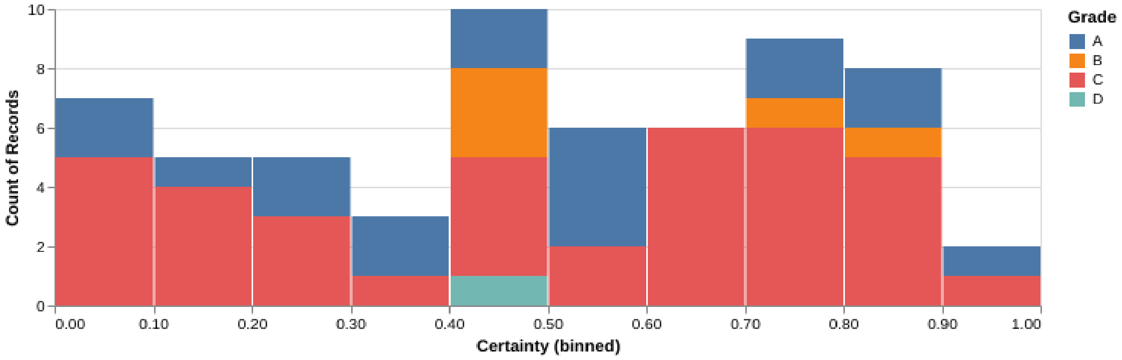


Figure 3: Histogram Count of the Model Certainty Colored by Grade

Our research identifies several significant potential improvements from incorporating our framework into the ISTRAP review process.

1. **Efficiency and Speed:** Manually checking every ISTRAP against dozens of detailed requirements is time-consuming. This software automates the initial review, drastically reducing the time needed per document. Procurement personnel can process more ISTRAPs faster.
2. **Granular Feedback:** Instead of a simple pass/fail, the system provides a grade for each *specific requirement*. This immediately highlights the exact sections or requirements within an ISTRAP that are weak (e.g., grades C, D, F) and need attention or revision.
3. **Focused Review:** Reviewers can use the results table to prioritize their efforts. They can quickly scan for low grades and focus detailed human reviews on those specific areas of non-compliance or weakness rather than re-reading sections where the AI is confidently graded as compliant (A or B; Reddy, 2024).
4. **Consistency:** The AI applies the same criteria (based on the system instruction and requirement definitions) to every document, reducing the variability and potential subjectivity that can occur with different human reviewers. This leads to more standardized initial compliance checks.
5. **Training and Template Improvement:** Common patterns of low grades can emerge by analyzing results across multiple ISTRAPs (if the code is extended to aggregate results from all documents). This data can identify systemic weaknesses in how ISTRAPs are being written, informing targeted training for personnel or improvements to ISTRAP templates and guidance.
6. **Risk Indication:** The “Certainty” score adds another layer of insight. A low grade with high certainty strongly indicates a problem. A low grade with low certainty might suggest the document is ambiguous or the AI struggled, warranting closer human inspection.
7. **Audit Trail:** The software execution and its results provide a documented record of the initial compliance check performed on each ISTRAP.

This framework provides an automated, granular, and consistent first-pass compliance check for **FAR, DFARS, and NMCARS**. It allows contracting officers to quickly identify potential issues, focus their review efforts efficiently, and visualize data in a way that leads to systemic improvements in **FAR, DFARS, and NMCARS** quality and compliance. The framework testing, while conducted specifically on the NMCARS Annex 18—ISTRAP requirements, establishes a foundational methodology applicable to structural compliance checks across all FAR, DFARS, and NMCARS regulations.

Conclusions

This research demonstrates the potential of leveraging generative AI, specifically NIPR GPT and RAG, to enhance the efficiency and accuracy of DoD acquisition package validation. The application of AI-driven tools shows promise in automating key tasks such as compliance checks, clause verification, and risk identification, ultimately reducing the administrative burden on contracting officers and minimizing the risk of errors.

Our findings indicate that AI models can effectively identify structural compliance gaps and potential omissions in contract documents, as evidenced by the ISTRAP and J&A analysis and the clause compliance assessment using PDS XML data. The integration of RAG further improves the accuracy and reliability of AI-generated outputs by providing access to real-time regulatory references, addressing limitations such as hallucinations and misinterpretations.

The development of a programmatic interface for batch summarization and ranking streamlines the analysis of multiple contract documents, enabling efficient identification of high-risk discrepancies and prioritization of review efforts. This automation not only accelerates the review process but also enhances consistency and objectivity in compliance assessments.

However, it is crucial to acknowledge the limitations of AI-driven solutions. The accuracy of AI models depends on the quality and completeness of the data they are trained on, and they may not always capture the nuances of regulatory language or adapt instantly to evolving regulations. Therefore, human oversight remains essential to validate AI-generated outputs and ensure comprehensive compliance.

Future research should focus on refining AI models through continuous learning and fine-tuning, integrating them with existing procurement systems, and developing best practices for human-AI collaboration. By addressing these challenges, the DoD can fully leverage the transformative potential of generative AI to modernize procurement practices, improve contract quality, and ultimately enhance mission readiness.

References

Atamel, M. (2025, January). *Evaluating RAG pipelines.* Medium. https://medium.com/google-cloud/evaluating-rag-pipelines-d99e007e625f

Entropy (information theory). (n.d.). In *Wikipedia*. Retrieved March 26, 2025, from https://en.wikipedia.org/wiki/Entropy_(information_theory)

Founding Minds. (2024, September 24). Chain-of-thought reasoning: The magic behind the o1 model. https://www.foundingminds.com/chain-of-thought-reasoning-the-magic-behind-the-o1-model/

Gadesha, V. & Kavlakoglu, E. (2024, August 12). *What is chain of thoughts (CoT)?*. IBM. https://www.ibm.com/think/topics/chain-of-thoughts

Google Cloud. (n.d.). *What is retrieval-augmented generation (RAG)?*. Retrieved March 26, 2025, from https://cloud.google.com/use-cases/retrieval-augmented-generation

Leblond, R., Alayrac, J.-B., Sifre, L., Pislar, M., Lespiau, J.-B., Antonoglou, I., Simonyan, K., Vinyals, O. (2021). *Machine translation decoding beyond beam search.* ArXiv. https://arxiv.org/pdf/2104.05336

Liu, L., Pan, Y., Li, X., & Chen, G. (2024). *Uncertainty estimation and quantification for LLMs: A simple supervised approach*. ArXiv. https://arxiv.org/html/2404.15993v1

Martyr, R. (2024, December 22). *Chain of thought prompting in AI: A comprehensive guide [2025]*. Orq. https://orq.ai/blog/what-is-chain-of-thought-prompting

Morgan, A. (2024, November 21). *Perplexity for LLM evaluation.* Comet. https://www.comet.com/site/blog/perplexity-for-llm-evaluation/

OpenAI. (2024, August 6). *Introducing structured outputs in the API*. https://openai.com/index/introducing-structured-outputs-in-the-api/

OUSD(A&S). (n.d.-a). *Data standards: Procurement data standard and other enterprise initiatives.* Retrieved April 2, 2025, from https://www.acq.osd.mil/asda/dpc/ce/ds/procurement-data-standard.html#:~:text=The%20Procurement%20Data%20Standard%20(PDS,including%20grants%20and%20other%20transactions.

OUSD(A&S). (n.d.-b). *Procurement data Standard and other enterprise initiatives.* Retrieved

April 2, 2025, from https://www.acq.osd.mil/asda/dpc/ce/ds/procurement-data-standard.html

Reddy, Y. (2024, October). *Avoiding LLM hallucinations and building LLM confidence scores*. Nanonets. https://nanonets.com/blog/how-to-tell-if-your-llm-is-hallucinating/

Shorten, C., Pierse, C., Smith, T. B., Cardenas, E., Sharma, A., Tengrove, J., & van Luijt B. (2024). *StructuredRAG: JSON response formatting with large language models.* ArXiv. https://arxiv.org/html/2408.11061v1

Tornetta, G. N. (2021). View of entropy methods for the confidence assessment of probabilistic classification models. *Statistica*. https://rivista-statistica.unibo.it/article/view/11479/13978

Villani, T. (2024, November 19). *How to implement chain-of-thought prompting for better AI reasoning.* New Jersey Innovation Institute. https://www.njii.com/2024/11/how-to-implement-chain-of-thought-prompting-for-better-ai-reasoning/

Zvornicanin, E. (2024, July, 16). *What are the evaluation metrics for RAGs?.* Baeldung on Computer Science. https://www.baeldung.com/cs/retrieval-augmented-generation-evaluate-metrics-performance

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

WWW.ACQUISITIONRESEARCH.NET