



EXCERPT FROM THE
PROCEEDINGS
OF THE
TWENTY-THIRD ANNUAL
ACQUISITION RESEARCH SYMPOSIUM AND
INNOVATION SUMMIT

VOLUME III
“ACCELERATING WARFIGHTING CAPABILITIES”

**Advancing Rule Development from NDAA Text through
Integrated LLM and Machine-Based Reasoning Tools**

Published: April 30, 2026

Approved for public release; distribution is unlimited.

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

Disclaimer: The views expressed are those of the author(s) and do not reflect the official policy or position of the Naval Postgraduate School, US Navy, Department of Defense, or the US government.



The research presented in this report was supported by the Acquisition Research Program, Graduate School of Defense Management at the Naval Postgraduate School.

To request defense acquisition research, please contact:

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
E: arp@nps.edu
www.acquisitionresearch.net

Copies of Symposium Proceedings and Presentations; and Acquisition Sponsored Faculty and Student Research Reports and Posters may be printed from the **NPS Defense Acquisition & Innovation Repository** at <https://dair.nps.edu/>.



ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE, AND MANPOWER
NAVAL POSTGRADUATE SCHOOL

Advancing Rule Development from NDAA Text through Integrated LLM and Machine-Based Reasoning Tools

Jose Ramirez-Marquez, PhD—is an Associate Professor and in the Department of Systems Engineering at Stevens Institute of Technology. His research focuses on the development of mathematical models for the analysis and computation of system operational effectiveness research and explores the interplay between data visualization and analytical decision-making. He conducted funded research for industry and government and published more than 100 refereed manuscripts in technical journals, book chapters, and industry reports. He holds degrees from Rutgers University in Industrial Engineering (PhD and MSc) and Statistics (MSc) and from Universidad Nacional Autonoma de Mexico in Actuarial Science. [jmarquez@stevens.edu]

Joshua Gorman—is a graduate research assistant at Stevens Institute of Technology, where he earned his Master's degree in Computer Science. In his current work, he focuses on building tools to analyze federal acquisition and defense policy, including systems that map NDAA provisions to FAR and DFARS regulations and interfaces for exploring various statutory sections. Previously, Gorman contributed to a range of applied research projects, including nuclear fallout visualization and the development of an LLM-powered dashboard for structured data extraction and analysis. [jgorman4@stevens.edu]

Akram Amer—is a graduate student at Stevens Institute of Technology completing his MEng in Applied Artificial Intelligence. He works as a graduate researcher at the Stevens AI Research Center where he has contributed to multiple DoD-funded projects involving NLP pipelines, LLM integration, and backend system development. [aamer@stevens.edu]

Douglas J. Buettner, PhD—is the AIRC Deputy Chief Scientist at the Stevens Institute of Technology and is an Adjunct Assistant Professor for the University of Utah's Systems Engineering Program. His doctoral research at the University of Southern California's Viterbi School of Engineering examined issues with DoD software-acquisition strategies, identifying acquisition optimizations. His research is found in Acquisition Games and Software Defects: Dynamics of the Software Developer's Dilemma. Dr. Buettner has BS and MS degrees in Physics, with a minor in Atmospheric Science from Oregon State University, and his PhD from USC is in Astronautical Engineering. [dbuettne@stevens.edu]

Brian Mayer—is a research scientist at the Sanghani Center. Mayer works with sponsors and collaborators to help apply Sanghani faculty research to solve government, corporate, and societal problems. He has helped deploy several long-term government-funded research projects to government sponsors. Mayer's research interests are in the areas of decision support, forecasting, intelligence, and human-guided systems. He received an MS and BS in Industrial and Systems Engineering from Virginia Tech. [bmayer@cs.vt.edu]

Nathan Self—is a research associate at the Sanghani Center. Self worked for a number of years in industry before joining Virginia Tech as a graduate assistant working on interactive visualizations for the EMBERS event forecasting project. He is now working on custom interactive visualizations for the Sanghani Center projects. He received a BS in Computer Science from the College of William and Mary in 2010 and an MS in Computer Science from Virginia Tech. [nwself@vt.edu]

Naren Ramakrishnan, PhD—is the Thomas L. Phillips Professor of Engineering at Virginia Tech and leads AI and machine learning for Virginia Tech's Innovation Campus. He directs the Sanghani Center for AI and Data Analytics and the Amazon-Virginia Tech Initiative in Efficient and Robust Machine Learning. Ramakrishnan's research interests span data science, forecasting, urban analytics, recommender systems, and computational epidemiology. He is a fellow of the Association for Computing Machinery (ACM), the American Association for the Advancement of Science (AAAS), and the Institute of Electrical and Electronics Engineers (IEEE). [naren@cs.vt.edu]

Harith Laxman—is a graduate student and Research Assistant at the Sanghani Center. He has previously worked on Reinforcement Learning (RL) and Conversational Recommendation Systems using text embeddings. His current work focuses on combining RL and LLMs for recommendation tasks. [harithlaxman@vt.edu]



Abstract

This research builds upon prior efforts by the authors to streamline Defense Federal Acquisition Regulation Supplement (DFARS) rule development from National Defense Authorization Act (NDAA) text using AI-based tools. The current research effort focuses on developing a unified, web-based interface that integrates previously developed prototypes and incorporates advanced Natural Language Processing (NLP), Large Language Models (LLMs), and Machine-Based Reasoning (MBR) techniques to improve automation in identifying, extracting, and recommending regulatory language changes.

The proposed unified system connects modules for document ingestion, keyword and context identification, text summarization, clustering, and visualization—through an integrated backend and user interface, which will enable DPCAP staff to move seamlessly from NDAA review to DFARS draft generation. The tool also proposes novel Machine-Based Reasoning techniques that leverage LLM models for updating proposed rule language and summarizing public comments from Regulations.gov. The combined framework is being deployed on a secure, sponsor-accessible server and will be evaluated against real DFARS updates in collaboration with DPCAP subject matter experts. This work aims to significantly reduce manual analysis time, enhance the traceability of regulatory updates, and strengthen the Department of Defense's capacity to apply AI responsibly in acquisition policy modernization.

Keywords: Large Language Models (LLMs); Natural Language Processing (NLP); Department of Defense (DoD); National Defense Authorization Act (NDAA); Federal Acquisition Regulation (FAR); Defense FAR Supplement (DFARS).

Introduction

The Defense Pricing, Contracting, and Acquisition Policy (DPCAP), Defense Acquisition Regulations System (DARS) Directorate in the Office of the Assistant Secretary of Defense (OSD) is responsible for implementing statutes, executive orders, and policies through the timely update of the Federal Acquisition Regulation (FAR) and Defense FAR Supplement (DFARS) and issuance of memoranda and guidance. Fundamentally, they enable operations through business systems and standards.

DPCAP is responsible for yearly updates to the DFARS based on changes in the National Defense Authorization Act (NDAA), Small Business Administration (SBA) rule changes, U.S. Department of Labor (USDOL) rule changes, or through executive orders (EOs). Reading through the changes made necessary by these multiple sources to complete required DFARS updates is labor intensive for DPCAP staff. It also requires knowledge of all the rules in the FAR/DFARS to ensure that changes are made appropriately, and references are made to the correct sections of the FAR/DFARS.

Artificial Intelligence (AI) is a powerful tool that can automate many tasks and support humans with others, but it has its limitations. Therefore, system development has to be a deliberate process that should be guided by policy and end use. Limitations can include bias, explainability, and trustworthiness (i.e., the well-known Large Language Model (LLM) hallucination problem). [2-4] Proper policy and implementation can limit bias, increase accuracy, and improve human effectiveness. When implementing AI solutions, it is important to understand these limitations and to create environments where AI systems and humans work in tandem to obtain the best results possible. For many tasks, the critical importance of human judgment means AI should serve as a complementary tool to improve human efficiency rather than as a standalone solution.

This project resulted in a tool that provides a cost-effective, scalable, semi-automated capability for managing regulatory updates, while ensuring long-term efficiency and adaptability. This paper shares the process of evaluating, integrating, and improving AI methods developed



to assist DPCAP subject matter experts (SMEs) in the FAR/DFARS change process into a unified system. SMEs are currently executing the tasks manually.

Prior work was completed to identify AI methods and processes to automate certain tasks in the FAR/DFARS change process and assist the DPCAP team (DPCAP FAR/FARS Change section). Evaluation of those methods was performed to determine effectiveness and impact (Evaluation of Methods section). All processes were then incorporated into a unified system and made available to the sponsor for testing (Proposed Unified System section). Evaluation results and feedback led to the improvement of the text generation process for FAR/DFARS changes (Improvements to Generation of FAR/DFARS Text Changes section). The team shares benefits, lessons learned, and future work in the conclusion (Conclusions and Future Work section)

DPCAP FAR/DFARS Change Process

DPCAP SMEs go through a 6-step process (shown in Figures 1 and 2) to identify necessary FAR/DFARS changes and make necessary updates to the FAR/DFARS.

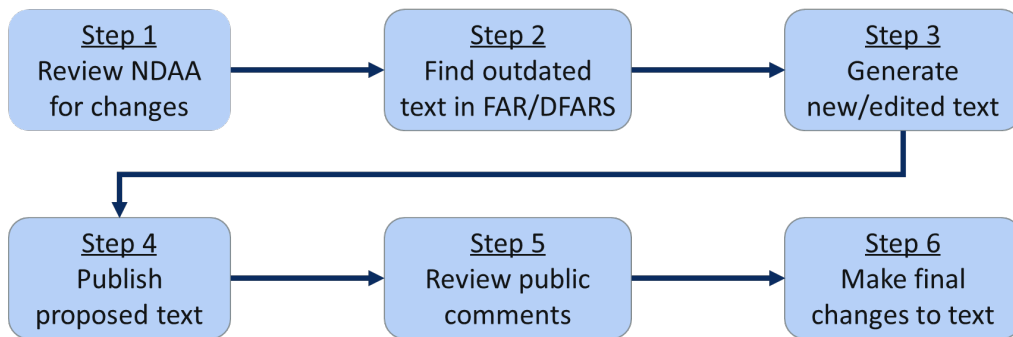


Figure 1. Task Flow Diagram



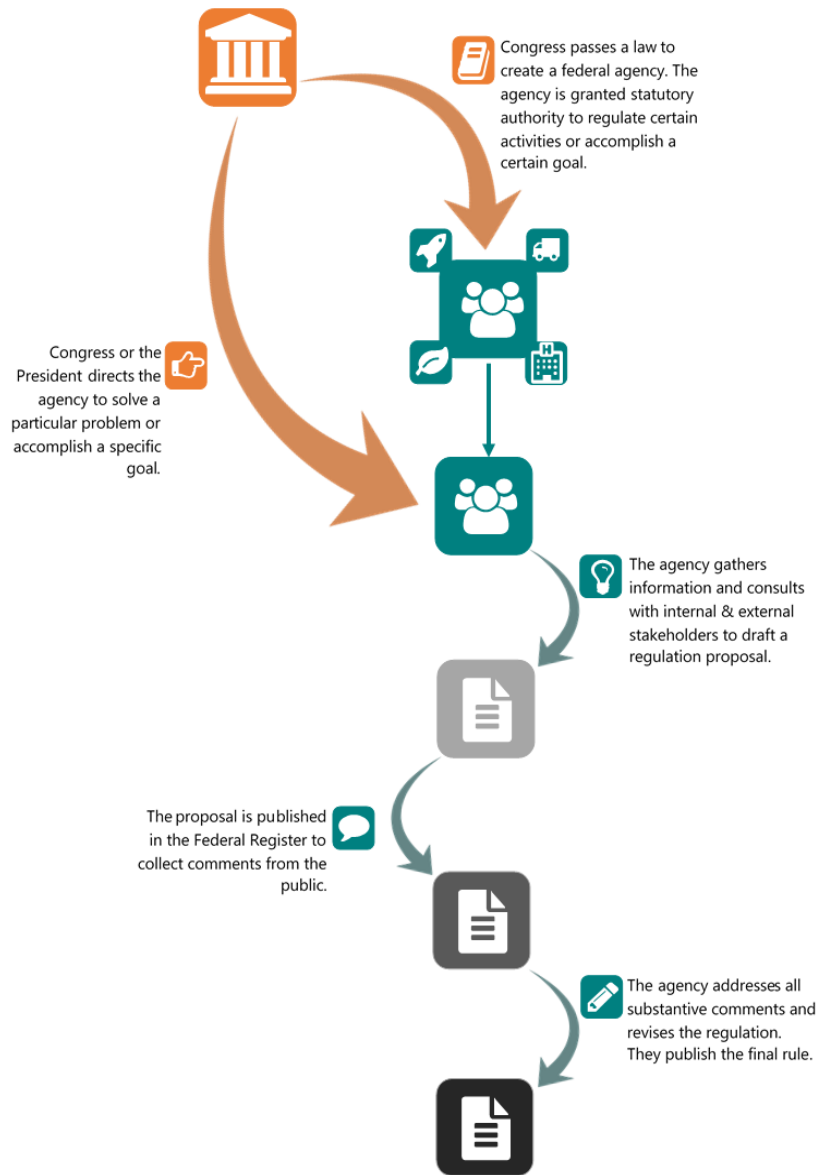


Figure 2. The Rule-making Process: How It Works

(<https://publiccommentproject.org/how-it-works>)

Step 1: NDAA Section Extraction and Storage

The AI methods developed support this process by converting NDAA, FAR, and DFARS PDFs into structured, searchable section records stored in the database and then identify NDAA sections that may require changes to FAR/DFARS.

The process for accomplishing this is as follows:

1. Upload a document for a specific year (NDAA year or FAR/DFARS regulation year).
2. The system parses the PDF and splits it into individual sections.
 - NDAA is split by section
 - FAR/DFARS are split by part → subpart → section
3. Each extracted section is stored in the database with key metadata, including:



- Section number
 - Title
 - Full text
 - Year
 - Document type (NDAA / FAR / DFARS)
 - For FAR/DFARS: part/subpart alignment
4. After the NDAA sections are stored, an analysis algorithm runs across all NDAA sections to identify which ones are likely to drive regulatory changes.
 5. These flagged sections are saved as NDAA candidate sections, allowing the system to reuse the same candidate list without reprocessing the NDAA.

A file-type detection and parsing pipeline was developed, enabling reliable text extraction from both the modern and archived legacy document formats that were incorporated to address older PDF formats used for DFARS/FAR documents prior to 2020. These changes required substantial updates to PDF parsing algorithms.

Step 2: NDAA ↔ FAR/DFARS Matching and Ranking

The AI methods we developed support this process by identifying which FAR/DFARS sections are most likely impacted by each NDAA change. This is accomplished by taking a specific NDAA year and FAR/DFARS regulation year. Then each NDAA candidate section is compared against all FAR and DFARS sections for that regulation year. A scoring algorithm evaluates relevance and potential impact. The top-ranked FAR/DFARS sections (e.g., top 10) are selected per NDAA section. These ranked matches are saved and made available to the user. This resulted in a structured, ranked matrix connecting each NDAA candidate section to its highest-scoring FAR and DFARS provisions for the selected regulation year. For each NDAA section, the system returns a prioritized list (e.g., top 10) of potentially impacted regulatory sections, forming a decision-support dataset for regulatory impact analysis.

Step 3: FAR/DFARS Text Changes

This process takes the selected NDAA year/section and its matched DFARS section as input then generates candidate text for a modified rule and the corresponding text to be published with a proposed rule in the Federal Register. The tool uses Chain of Thought (CoT) prompting [5] and requests specific steps (defined by the research team) and outputs to be performed to accomplish the task. Documents may not be formatted to FAR/DFARS guidelines. The output of the tool (i.e., LLM) needs to be edited by an SME but can serve as a starting point. Reference and identification numbers will often have placeholders. The tool can only modify current DFARS sections. It cannot draft new sections.

Step 4: Publish Proposed Text

This step was not a part of the AI automation effort due to external requirements.

Step 5: Public Comment Review

Comments for previously published DFARS proposals are downloaded from Regulations.gov. A user can then select the proposed regulation they would like to review. Duplicate comments are removed, and similar comments (based on semantic meaning) are grouped together. The groups are then summarized by an LLM (in this case Generative Pre-trained Transformer [GPT] 4o as a representative of the available Non-classified Internet Protocol Router GPT [NIPRGPT] [6]). If there are sufficient comments, a subset of the group's comments are used for the summarization to reduce LLM token requests without drastically



affecting the summary since comments have been grouped based on similarity. Finally, the tool provides an overarching summary for the entire set of comments.

Step 6: Use Comments to Make Changes to the Proposed Text

For each group of comments, the tool provides edit recommendations to the DFARS section based on what the comments are suggesting. The user must choose whether to incorporate the suggestions from these comments or not.

The research team also investigated providing these suggestions to the LLM along with the full text from the proposed rule, but the resulting text was often too long and often longer than the LLM's token limits would allow. The Task 3 pipeline could instead be repurposed for generating these modifications.

Evaluation of Methods

NDAAs Section Extraction

The system processed NDAs from 2010–2025. This resulted in a normalized, year-specific repository of NDAs stored as structured records, along with a reproducible set of NDA candidate sections identified through automated statutory impact analysis. These candidate sections represent provisions likely to require FAR or DFARS regulatory updates and serve as the controlled input set for downstream regulatory mapping. A corresponding quality index was also produced. The quality index was calculated for 2010–2023. Discrepancies remain, however, primarily stemming from parsing limitations with the older PDFs, which are being addressed through an ongoing refinement effort of the extraction methods. The index is obtained as:

$$Q(t) = \frac{E_{\text{ext}}(t)}{E_{\text{exp}}(t)} \quad (1)$$

- $E_{\text{ext}}(t)$ is the number of extracted items for NDA year t ,
- $E_{\text{exp}}(t)$ is the number of expected items for NDA year t .

For example, for Fiscal Year (FY) 2020, the candidate extraction process produced the following NDA sections:

[2821, 1648, 1212, 1123, 890, 888, 885, 874, 873, 872, 870, 865, 864, 861, 854, 853, 852, 849, 848, 847, 826, 825, 823, 822, 818, 816, 807, 803, 372, 322, 168]

These sections constitute the automated system's identified "golden" set for FY2020 and are passed as inputs into the regulatory matching algorithm for FAR/DFARS linkage analysis.



Table 1 Quality Index for Expected NDAA Sections Versus Extracted NDAA Sections

Fiscal Year	Extracted	Expected	Quality Index	Missed Section Numbers
2010	15	16	94%	745
2011	18	20	90%	834, 893
2012	26	28	93%	841, 834
2013	18	20	90%	802, 1651
2014	7	7	100%	
2015	15	16	94%	745
2016	25	26	96%	1214
2017	41	43	95%	873, 1212
2018	30	31	97%	1214
2019	30	31	97%	888
2020	31	31	100%	
2021	28	29	97%	832
2022	20	20	100%	
2023	19	19	100%	

Note that ground-truth “golden” sections are derived from manual analyst identification of NDAA provisions likely to require FAR or DFARS regulatory updates. The automated extraction process is evaluated against this benchmark. “Expected” denotes the analyst-identified count, while “Extracted” reflects the number successfully identified by the system. The Quality Index represents extraction accuracy relative to the validated benchmark. “Missed Sections” lists analyst-identified sections that were not captured by the automated method.

NDAA ↔ FAR/DFARS Matching

This section defines a second set of metrics used to evaluate system performance in identifying FAR/DFARS mappings associated with NDAA sections for a given fiscal year. “Sections with Published FAR/DFARS Mappings” represents the number of NDAA sections within a given fiscal year that have validated FAR/DFARS mappings and therefore serve as ground-truth references. “Sections < 50% Mapping Coverage” identifies sections for which the system retrieved fewer than 50% of the known FAR/DFARS mappings. The associated “Quality Index (%)” reflects overall mapping coverage performance relative to validated ground-truth associations for that fiscal year.

For clarity, the metrics can be described as follows:

Let S_y denote the set of NDAA sections in fiscal year y . For each section $s \in S_y$, define:

$G(s)$: the set of validated ground-truth FAR/DFARS mappings associated with section s

$R(s)$: the set of FAR/DFARS mappings retrieved by the system for section s



Evaluation is restricted to sections with at least one validated ground-truth mapping.

Sections with Published FAR/DFARS Mappings

The subset of sections with validated FAR/DFARS mappings is defined as:

$$S_y^G = \{s \in S_y: |G(s)| > 0\} \quad (2)$$

The total number of such sections is:

$$N_y^G = |S_y^G| \quad (3)$$

Thus, “Sections with Published FAR/DFARS Mappings” represents the number of NDAA sections in fiscal year y for which validated FAR/DFARS mappings exist and can therefore serve as ground-truth references.

Section-Level Coverage

For each section s , mapping coverage is defined as:

$$C(s) = \frac{|R(s) \cap G(s)|}{|G(s)|} \quad (4)$$

This quantity measures the proportion of validated mappings for section s that were successfully retrieved by the system.

Sections with Less Than 50% Mapping Coverage

The subset of sections for which system retrieval covers fewer than half of the validated mappings is defined as:

$$S_y^{<50} = \{s \in S_y^G: C(s) < 0.5\} \quad (5)$$

The number of such sections is:

$$N_y^{<50} = |S_y^{<50}| \quad (5)$$

Accordingly, “Sections < 50% Mapping Coverage” identifies the number of NDAA sections with retrieved FAR/DFARS mappings that account for less than half of the known validated mappings.

Quality Index (%)

The Quality Index reflects overall mapping coverage performance relative to validated ground-truth associations for fiscal year y . Two formulations may be used depending on the intended interpretation.

Micro-Averaged Quality Index

The micro-averaged version weights sections according to the number of validated mappings they contain:

$$QI_y^{\text{micro}} = \frac{\sum_{s \in S_y^G} |R(s) \cap G(s)|}{\sum_{s \in S_y^G} |G(s)|} \times 100 \quad (7)$$

This formulation measures the percentage of all validated mappings in fiscal year y that were correctly retrieved by the system.

Interpretation

The metrics may be interpreted as follows:



N_y^G : number of NDAA sections in fiscal year y with published and validated FAR/DFARS mappings

$N_y^{<50}$: number of sections for which the system retrieved fewer than 50% of the known mappings

QI_y^{micro} overall percentage of validated FAR/DFARS mappings captured by the system

Table 2. Quality Index for NDAA Section to FAR/DFAR Section(s) Mappings

Fiscal Year	Sections w/ Published FAR/DFARS Mappings	Sections <50% Mapping Coverage	Quality Index (%)
2010	15	2	87%
2011	20	3	85%
2012	31	7	77%
2013	24	7	71%
2014	7	1	86%
2015	18	7	61%
2016	27	6	78%
2017	50	16	68%
2018	33	11	67%
2019	28	9	68%
2020	31	5	84%
2021	29	9	69%
2022	20	6	70%
2023	19	5	74%

Evaluation of FAR/DFARS Text Changes

The evaluation of the proposed text changes was performed using traditional metrics such as Bilingual Evaluation Understudy (BLEU) [7] and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [8], as well as the more recent and robust method of LLM-as-a-Judge [9]. Finalized section changes were used as ground truth samples. The LLM judge was asked to grade the following factors (out of five):

- Change Identification
- Content Accuracy
- Completeness
- Structural Fidelity (basically if the ordering of the points are same or not)

The model was provided with a 1:1 mapping of NDAA section to DFARS section and then asked to create the proposed modified DFARS section. The generated section was compared using the above methods to the ground truth sample. There were 172 DFARS cases with ground truth samples from 2010–2024 that were used for evaluation.



In the case where there was a Many:1 or Many:Many relationship between NDAA section and DFARS section we simplified it by taking each 1:1 pair and passing it through the text generator and evaluating the results against the ground truth (1:Many instances are not a problem because each NDAA:DFARS relationship can be represented with a 1:1 resulting in a DFARS changed by a single NDAA). In these instances, the text generator did not have all of the context that the ground truth had so it should not be expected to be perfect, but this was the selected method for evaluating the changes without complicating the correlation of changes to sections. Additionally, the tool currently can only act (generate) on 1:1 changes so it was best to evaluate what the tool could do.

Table 3. LLM Judged Quality of System Generated DFARS Text

Dimension	Average
LLM-Judge: Change Identification	2.26
LLM-Judge: Content Accuracy	2.00
LLM-Judge: Completeness	2.33
LLM-Judge: Structural Fidelity	2.20
BLEU (Max 85.4, Min 0)	28.8

Proposed Unified System

The objective was to create a unified system (backend and user interface) which enables DPCAP staff to move seamlessly from NDAA review to DFARS draft generation.

The developed system includes a backend and a publicly accessible interface available to the sponsor. The modules described in Section 2 (document ingestion, keyword and context identification, text summarization, clustering, and visualization) are connected in an integrated backend that drives the user interface.

The system:

- Ingests NDAAs, FAR, and DFARS
- Extracts NDAA sections
- Finds matching FAR/DFARS sections for each extracted NDAA section
- Generates text changes to the FAR/DFARS and a change summary for the Federal Register.
- Ingests, clusters, and summarizes public comments for all proposed FAR/DFARS changes

The backend steps are completed and then the front-end imports the necessary information for the interface. This includes:

- all sections of interest from the NDAA (a link is provided to the full NDAA)
- sections of the DFARS matching the identified NDAA sections, mentioned above (a link is provided to the full DFARS)
- generated DFARS proposed rules and Federal Register text for each NDAA-DFARS pair
- comments from selected proposed DFARS rules



- groups of comments along with summaries and proposed revisions

The front-end interface displays all the information to support the user as they work through the steps in the pipeline.

1. Review NDAA for changes
2. Find outdated text in FAR/DFARS
3. Generate new/edited text
4. Publish proposed text (not in current scope)
5. Review public comments
6. Make final changes to text

Virginia Tech provisioned a virtual machine to serve that runs the backend code and is the web server for the front-end code. The front-end is publicly hosted behind a secure login to make it accessible to the sponsor for testing.

The system was made available to DPCAP staff for review and assessment. Additionally, a video demonstration and visual instruction manual were developed to help users navigate the tool. Screenshots of using the system to go through the DFARS change process are provided in Figure 5 through Figure 10.

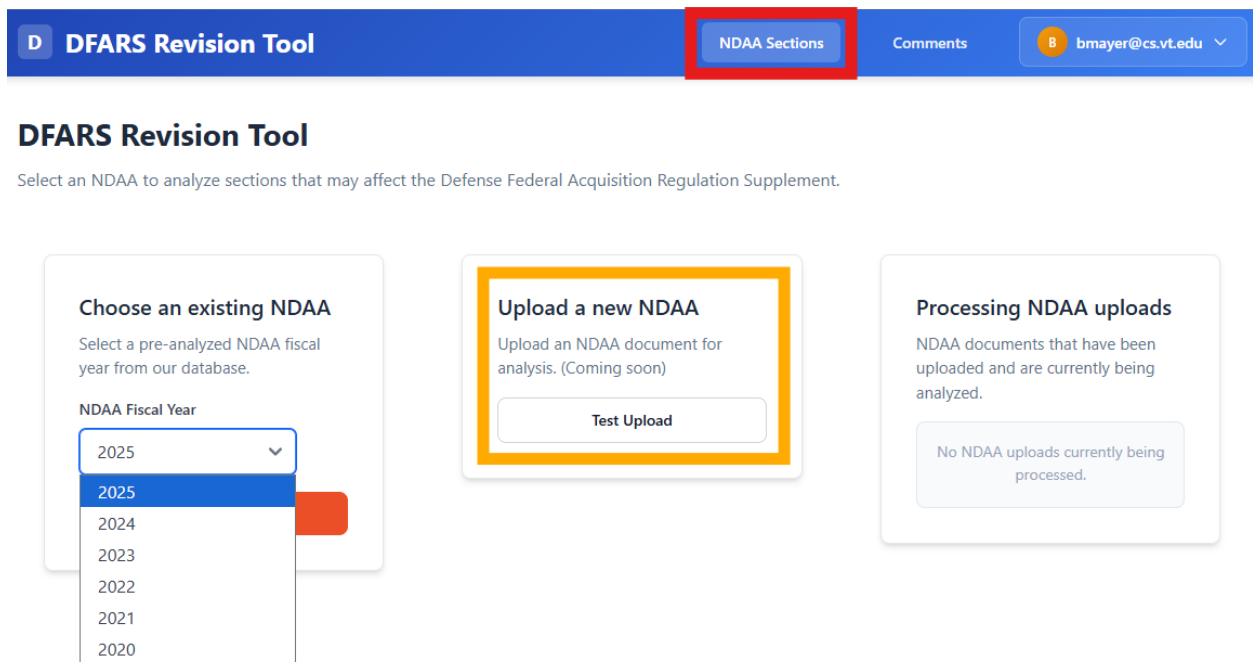


Figure 3. Selecting or Uploading an NDAA

D DFARS Revision Tool NDAAs Sections Comments B bmayer@cs.vt.edu

Choose NDAAs / NDAAs FY 2025 Sections

NDAAs Sections for Fiscal Year 2025

Sections that may require changes to the DFARS.

NDAAs Fiscal Year: December 23, 2024 Open NDAAs

DFARS Effective Date: October 1, 2025 Open DFARS

Our algorithm identified 30 NDAAs sections that may incur changes to the DFARS

NDAAs SECTION 335

AUTHORITY FOR GOVERNMENT-OWNED, GOVERNMENT-OPERATED FACILITIES TO ACCESS PRODUCTION BASE SUPPORT FUNDS.

Show DFARS matches →

Section SEC. 335.: AUTHORITY FOR GOVERNMENT-OWNED, GOVERNMENT-OPERATED FACILITIES TO ACCESS PRODUCTION BASE SUPPORT FUNDS.

Note:
 Deadline.
 Regulations.
[10 USC note](#)
 prec. 7532.

Not later than 180 days after the date of the enactment of this Act, the Secretary of Defense shall prescribe regulations providing that Government-owned, Government-operated facilities are eligible to receive production base support funding from the Army.

Figure 4. Reviewing the NDAAs and Identified Sections

D DFARS Revision Tool NDAAs Sections Comments B bmayer@cs.vt.edu

NDAAs SECTION 802

LIMITATION ON CERTAIN OPTIONS FOR COST CONTRACTS.

Show DFARS matches →

Section SEC. 802.: LIMITATION ON CERTAIN OPTIONS FOR COST CONTRACTS.

(a) AMENDMENTS.—
 Section 3322 of title 10, United States Code, is amended by adding at the end the following new subsection:

“(d) LIMITATION ON CERTAIN OPTIONS.—

Figure 5. Selecting an NDAAs Section to Find DFARS Matches For



DFARS Revision Tool NDAA Sections Comments bmayer@cs.vt.edu

Choose NDAA / NDAA FY 2025 / Section 802 – DFARS matches

NDAA SECTION 802
LIMITATION ON CERTAIN OPTIONS FOR COST CONTRACTS.

[Read Section Text →](#)

Matched DFARS Sections
 The following DFARS sections have been matched to this NDAA section. Click "Generate suggested rule" to view the AI-generated proposed rule text.

Part 217 - Special Contracting Methods
 Subpart 217.1 - Multiyear Contracting

Section 217.700 Scope of subpart.

217.700 Scope of subpart.
 This subpart—
 (a) Implements section 854 of the National Defense Authorization Act for Fiscal Year 2005 (Pub. L. 108-375), section 801 of the National Defense Authorization Act for Fiscal Year 2008 (Pub. L. 110-181), and section 806 of the National Defense Authorization Act for Fiscal Year 2010 (Pub. L. 111-84); and
 (b) Prescribes policy for the acquisition of supplies and services through the use of contracts or orders issued by non-DoD agencies.

[Show generated changes →](#)

Part 208 - Required Sources Of Supplies And Services
 Subpart 208.4 - Federal Supply Schedules

Section 208.705 Procedures.

[Show generated changes →](#)

Figure 6. Reviewing DFARS Matches and Selecting a Match to Generate a Modification For

DFARS Revision Tool NDAA Sections Comments bmayer@cs.vt.edu

Choose NDAA / NDAA FY 2025 / Section 802 / Suggested rule – DFARS 217.7402

NDAA Section 802 → DFARS 217.7402
Suggested Rule: 217.700 Scope of subpart.
 Part 217 - Special Contracting Methods / Subpart 217.1 - Multiyear Contracting

Text of DFARS 217.7402	Proposed Change	Federal Register Summary
<p>217.7402 Limitation on Certain Options for Cost Contracts</p> <p>Scope This clause applies to all cost reimbursement contracts for the development of a major system, as defined in 10 U.S.C. 3041, awarded by the Department of Defense. It specifically governs contracts involving fixed-priced options for production quantities and limits the number of low-rate initial production (LRIP) lots to no more than one. This limitation applies regardless of contract value or contractor classification. Waivers to this limitation may be granted only by the service acquisition executive or, in the case of joint programs, the Secretary of Defense.</p> <p>Definitions For purposes of this clause: (1) Covered contract means a cost reimbursement contract for the development of a major system. (2) Low-rate initial production (LRIP) has the meaning given in 10 U.S.C. 4231. (3) Major system has the meaning provided in 10 U.S.C. 3041.</p> <p>Restrictions (1) Except as provided in paragraph (2), the contract shall limit the number of low-rate initial production lots procured using fixed-priced options to not more than one. (2) A waiver to the limitation in paragraph (1) may be granted only under the following conditions: (i) The waiver must be formally determined by the service acquisition executive of the military department concerned or, for joint programs, by the</p>		

Figure 7. Reviewing Generated Text for the Modification



The screenshot shows the 'DFARS Revision Tool' interface. At the top, there are navigation elements: 'D DFARS Revision Tool', 'NDAA Sections', 'Comments', and a user profile 'bmayer@cs.vt.edu'. The main heading is 'Comment Summaries for Proposed Rules' with a sub-heading 'Analysis of public comment on proposed rules.' On the left, a 'Search Proposals' sidebar lists several FAR cases. The case 'FAR-2023-0021-0001 - Acquisition Regulation: Pay Equity and Transparency in Federal Contracting' is highlighted with a yellow box. The main content area displays the title 'Acquisition Regulation: Pay Equity and Transparency in Federal Contracting' and a 'Comment Summary (based on 25 comments)'. The summary text discusses concerns about legal compliance and the definition of 'performance in the United States'. Below the summary is a 'Comment Groups' section with a dropdown menu showing 'Legal Concerns Over 41 U.S.C 6707(c)', 'Clarification on USAID Contract Performance', 'File Attachment Discussion Highlights', and 'Support for FAR Case 2023-021: Pay Equity'.

Figure 8. Selecting a DFARS Proposal and Reviewing the Summary of Public Comments

This screenshot provides a more detailed view of the 'DFARS Revision Tool'. The 'Search Proposals' sidebar on the left is expanded, showing a search bar and a list of FAR cases. The selected case, 'FAR-2023-0021-0001 - Acquisition Regulation: Pay Equity and Transparency in Federal Contracting', is highlighted in blue. The main content area shows the 'Acquisition Regulation: Pay Equity and Transparency in Federal Contracting' title and a 'Comment Summary (based on 25 comments)'. Below the summary is the 'Comment Groups' section, where the 'Legal Concerns Over 41 U.S.C 6707(c)' group is expanded. This group contains a 'Summary' section (highlighted with a red box) stating that concerns have been raised regarding potential violations of 41 U.S.C 6707(c) and that commenters are questioning the legality and compliance of the government's actions. Below the summary is a 'Revision Suggestion' section (highlighted with a yellow box) that advises revising the proposed rule to include a clear compliance section, providing specific language and examples of actions to mitigate potential violations. At the bottom of the comment group, there is a 'View All Comments' button (highlighted with a blue box).

Figure 9. Reviewing the Summary of a Comment Group, the Proposed Revision Based on the Comments, and Viewing the Raw Text from All the Grouped Comments



Improvements to Generation of FAR/DFARS Text Changes

The text generation process was improved by using an LLM (GPT 4.1) agent to evaluate the need for changes then another to collect additional context which is fed into a third LLM agent to determine the necessary changes given the input sections and the identified context. Finally, a fourth LLM agent is provided with FAR drafting guidelines and the necessary changes to generate the modified section text. The process is depicted in Figure 10.

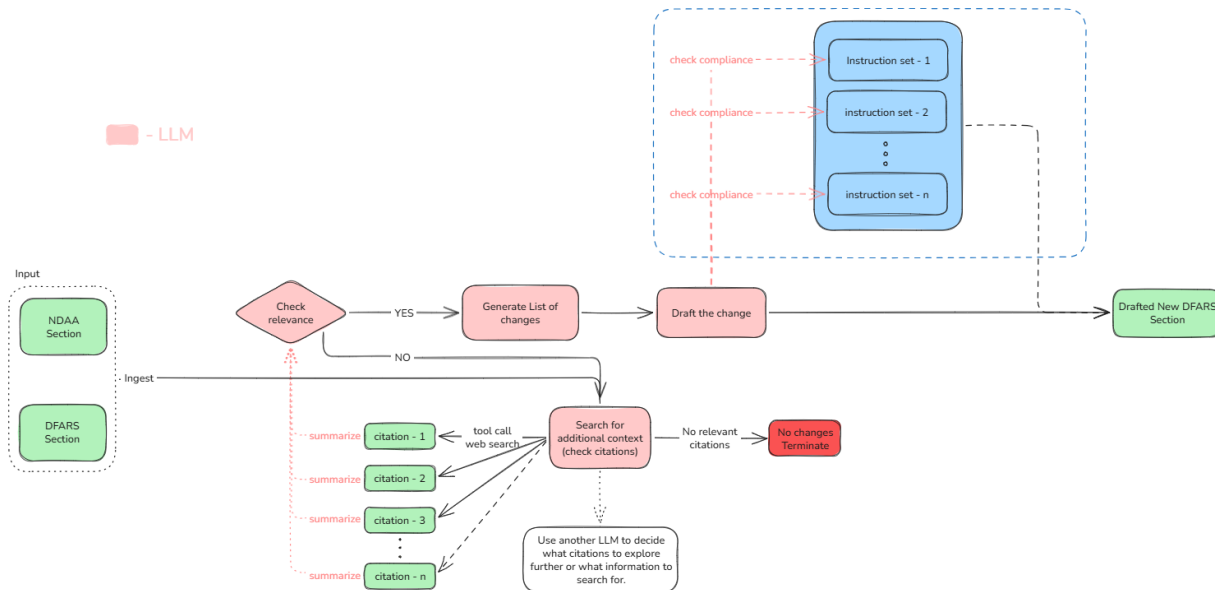


Figure 10. Process for Generating Changes to a DFARS Section Given an Old DFARS Section and an NDAAs Section

The first agent is provided the NDAAs section and DFARS section. It is tasked with identifying references to other documents (e.g., U.S. Code, NDAAs, FAR, DFARS) and building up a knowledge base to support the other steps. Once this context is identified the second agent is provided with the NDAAs section the DFARS section and all the identified context documents. The second agent is asked to determine if a change is necessary to the DFARS given the NDAAs section and all the context.

If no change is necessary, the user is informed that no text was generated because the system could not identify an appropriate change. If a change is appropriate, the third agent is tasked with describing the necessary changes. This description is then provided to the fourth and final agent to generate the modified text.

There are a few significant limitations to this process. First, the process currently only supports modifications to sections. To account for cases where a completely new DFARS section is required, the pipeline will have to be fed the entire DFARS subpart. This would need to be determined by the second agent, and provided to the third agent.

Additionally, while the first agent does identify the related citations and provide them to the other agents, this tool is using generative AI and it cannot be relied upon to make accurate citations [10]. The system may generate text with citations but there is no guarantee (and there should be no expectation) that these citations will be accurate

Finally, as previously stated, the system currently only handles a 1:1 mapping of NDAAs sections to DFARS sections. In many cases, multiple NDAAs sections may change 1 or more DFARS sections. This would require identifying all the NDAAs sections at once which is not



currently a capability of the NDAA-DFARS matching process. User intervention is necessary to review proposed text and therefore they can incorporate changes from multiple NDAA sections at once using the text changes from each 1:1 mapping automatically generated from the system.

Despite some limitations, the model still provides a reasonable draft (as seen in the results in the Evaluation of Methods section), but not without requiring SME review and potential edits.

Additionally, the multi-agent process forces the agents to perform tasks with more reasoning, which consequently provides additional explainability to the user. This explainability could be further leveraged by the rest of the system and interface for further improvements and adoption.

Conclusions and Future Work

Conclusions

This research demonstrates the feasibility and value of integrating LLMs, Natural Language Processing (NLP), and Machine-Based Reasoning (MBR) into a unified framework to support the DFARS rule development process. By further evaluating, refining, and consolidating previously independent AI-driven capabilities into a single, end-to-end system, the project enables a more streamlined transition from NDAA analysis to DFARS rule drafting and revision.

The evaluation results indicate that automated NDAA section extraction and FAR/DFARS mapping achieve consistently high levels of accuracy, providing a reliable foundation for downstream tasks. While variability remains in mapping coverage and text generation quality, particularly in complex or many-to-many relationships, the system has the potential for meaningful reduction in manual workload and acceleration of regulatory analysis. The incorporation of multi-agent LLM pipelines further enhances contextual understanding and improves the quality of generated regulatory text, though human SME oversight remains essential to ensure correctness, compliance, and trustworthiness. Additionally, the integration of public comment analysis introduces a scalable mechanism for incorporating stakeholder feedback into rule development, further strengthening the policy life cycle.

Despite these advances, the tool does have limitations. Current capabilities are constrained to one-to-one mappings between NDAA and DFARS sections, and the system does not yet fully support the creation of entirely new regulatory sections. Furthermore, challenges related to LLM reliability, including citation accuracy and potential hallucinations, necessitate continued emphasis on human-in-the-loop validation and robust governance frameworks. Additional fine-tuning is also imperative given the limited amount of ground truth data and risk of over-fitting.

Suggested Future Research and Next Steps

Future work should address the mentioned limitations (e.g., many:1 mappings, LLM reliability, model testing, and fine-tuning).

Future work should also focus on leveraging the LLM-as-a-judge (first agent from the text generation pipeline) to act as an additional tool in the NDAA identification and DFARS matching processes. In this scenario the matching models could be tuned to increase recall, and the LLM-as-a-judge could be used to increase precision. This could result in higher recall and higher precision of the joint system. The LLM agents' output could also be further leveraged to increase traceability between regulatory documents and statutory mandates, providing further explainability to the user.



Final Thoughts

Overall, this work highlights the transformative potential of AI-assisted systems in regulatory environments, offering a scalable, semi-automated approach that augments human expertise while maintaining the rigor required for defense acquisition policy development.

Acknowledgements

We would like to thank Ms. Dawn Messer and Ms. Kimberly Ziegler, OSD Defense Pricing, Contracting and Acquisition Policy, for their sponsorship and support of this analytical effort.

The views, findings, and conclusions in this document are solely those of the authors and do not necessarily reflect the views or positions of the United States Government (including the DoD or any other government personnel) or the Stevens Institute of Technology.

References

- Under Secretary of Defense for Acquisition and Sustainment. (2011, April 30) *FAR drafting guide*. Department of Defense.
https://www.acq.osd.mil/dpap/dars/docs/far_dfars_guide/FAR%20Drafting%20Guide--April%2030,%202011.pdf
- Alansari, A., & Luqman, H. (2026). *Large language models hallucination: A comprehensive survey*. ArXiv. [arXiv:2510.06265v3](https://arxiv.org/abs/2510.06265v3)
- Anh-Hoang, D., Tran, V., Nguyen, L. M. (2025). *Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior*. Front Artif Intell. <https://doi.org/10.3389/frai.2025.1622292>. PMID: 41098969; PMCID: PMC12518350.
- Tauman Kalai, A., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why language models hallucinate*. ArXiv. [arXiv:2509.04664v1](https://arxiv.org/abs/2509.04664v1)
- Wei, J. et al. (2023, January 10). *Chain-of-thought prompting elicits reasoning in large language models*. ArXiv. <https://doi.org/10.48550/arXiv.2201.11903>.
- Secretary of the Air Force Public Affairs. (2024 Jun. 10) *Department of the Air Force launches NIPRGPT*. <https://www.af.mil/News/Article-Display/Article/3800809/department-of-the-air-force-launches-niprgpt/>
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).
- Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81).
- Gu, J., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., ... & Guo, J. (2024). A survey on LLM-as-a-judge. *The Innovation*.
- Buchmann, J., & Gurevych, I. (2025). Citation failure: Definition, analysis and efficient mitigation. ArXiv. *arXiv preprint arXiv:2510.20303*.





ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF ACQUISITION, FINANCE, AND MANPOWER
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

WWW.ACQUISITIONRESEARCH.NET