NPS-AM-26-221

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Enhancing Decision Accuracy in DoD Acquisition: Integrating Artificial Intelligence with Reference Class Forecasting

June 2026

MCPO Monte L. Ellis Jr., USN

Thesis Advisors: Jeffrey R. Dunlap, Lecturer
Dr. Christina C. Hart, Faculty Associate

Department of Acquisition, Finance and Manpower

Naval Postgraduate School

Prepared for the Naval Postgraduate School, Monterey, CA 93943

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

The research presented in this report was supported by the Acquisition Research Program of the Department of Acquisition, Finance and Manpower at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact the Acquisition Research Program (ARP) via email, arp@nps.edu or at 831-656-3793

Acquisition Research Program
Department of Acquisition, Finance and Manpower
Naval Postgraduate School

ABSTRACT

This capstone examines how integrating artificial intelligence (AI) with reference class forecasting (RCF) can improve decision accuracy in Department of Defense (DoD) acquisition. Persistent cost overruns and schedule delays, driven by optimism bias and planning fallacy, highlight the limits of traditional forecasting. These shortfalls routinely undermine mission readiness and erode fiscal discipline. While RCF enhances accuracy by anchoring estimates in historical data, its use in the DoD is limited by scalability and data-access challenges.

This study uses a qualitative design combining policy review, comparative case analysis, and conceptual modeling. The findings indicate that AI can support the automation of reference-class construction from unstructured acquisition data and enable probabilistic forecasting, improving cost and schedule realism. Supported by prior literature and simulated analysis, the results also suggest that this approach can strengthen technology readiness assessments (TRA) by incorporating risk-based probability bands, thereby highlighting the value of probabilistic evidence in early acquisition decision-making.

Recommendations include phased AI–RCF implementation, governance standards for transparency, and integration into milestone artifacts like TRAs and Life cycle Sustainment Plans. Institutionalizing this approach would embed empirical rigor into acquisition decisions, reduce systemic risk, and advance the DoD's shift toward data-driven reform.

THIS PAGE INTENTIONALLY LEFT BLANK

ABOUT THE AUTHOR

Master Chief Monte L. Ellis Jr. enlisted in the Navy in 1996 and has served in diverse aviation maintenance and senior leadership roles across the fleet. He qualified on multiple aircraft platforms—including the TA-4J, UH-3H, H-60B/F/H, P-3 Orion, E-6B Mercury, and E-2C Hawkeye—and deployed aboard several carriers and destroyers while earning both EAWS and ESWS qualifications. His assignments include tours with VC-8, FRC SEAOPDET at NAS Oceana, FRC Patuxent River, HSL-51, HS-5, COMSTRATWING ONE, VP-1, the Senior Enlisted Academy as a Faculty Advisor, and VAW-116. He holds a degree in Electronic Engineering Technology from Grantham University and a Bachelor of Science in Organizational Leadership from Penn State, and he currently serves as the Senior Enlisted Advisor at NAWCTSD while completing his master's degree at the Naval Postgraduate School. His personal awards include four Navy Commendation Medals and two Navy Achievement Medals, and he and his wife Kimberli have five children.

THIS PAGE INTENTIONALLY LEFT BLANK

NPS-AM-26-221

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Enhancing Decision Accuracy in DoD Acquisition: Integrating Artificial Intelligence with Reference Class Forecasting

June 2026

MCPO Monte L. Ellis Jr., USN

Thesis Advisors: Jeffrey R. Dunlap, Lecturer
Dr. Christina C. Hart, Faculty Associate

Department of Acquisition, Finance and Manpower

Naval Postgraduate School

Prepared for the Naval Postgraduate School, Monterey, CA 93943

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AAF	adaptive acquisition framework
AHE	Advanced Hawkeye
AI	artificial intelligence
AoA	analyses of alternatives
APUC	average procurement unit cost
CADE	Cost Assessment Data Enterprise
CAPE	Cost Assessment and Program Evaluation
CER	cost estimating relationship
CTE	critical technology elements
DAES	Defense Acquisition Executive Summary
DAU	Defense Acquisition University
DCAPE	Director of Cost Assessment and Program Evaluation
DoD	Department of Defense
EMALS	Electromagnetic Aircraft Launch System
EVAMOSC	Enterprise Visibility and Management of Operating and Support Costs
GAO	Government Accountability Office
G/ATOR	Ground/Air Task Oriented Radar
HASC	House Armed Services Committee
ICG	Internal Consulting Group
IOC	initial operational capability
ITEP	Improved Turbine Engine Program
ITRA	independent technology readiness assessment
JCIDS	Joint Capabilities Integration and Development System
JLTV	Joint Light Tactical Vehicle
JSF	Joint Strike Fighter
LCS	Littoral Combat Ship
LCSP	life-cycle sustainment plan
MAE	mean absolute error
MDAP	major defense acquisition program
MIDS	Multifunctional Information Distribution System

ML	machine learning
MSAR	Modernized Selected Acquisition Report
NLP	natural language processing
O&S	operating and support
PAUC	program acquisition unit cost
POM	Program Objective Memorandum
PSM	product support manager
PPBE	Programming, Planning, Budgeting, and Execution
RCF	reference class forecasting
RMP	Radar Modernization Program
RMSE	root mean square error
SAR	Selected Acquisition Report
SHAP	Shapley Additive Explanation
SME	subject matter expert
TRA	technology readiness assessments
TRL	technology readiness level
UK	United Kingdom
WBS	work breakdown structure

DISCLOSURES

Generative artificial intelligence tools were used in a limited, support-only capacity during the research and writing process, with the knowledge and permission of the thesis advisory team. Specifically, Microsoft Copilot was used to assist with document retrieval, text parsing, visualization, and figure generation. All analytical judgments, modeling logic, data interpretation, conclusions, and recommendations were developed solely by the author. No generative artificial intelligence tools were used to generate original research findings or replace author analysis.

All sources of data used in this thesis were publicly available or otherwise authorized for academic use. No proprietary, restricted, or classified information was accessed or included.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

This chapter outlines the purpose and scope of the capstone project and introduces the research questions guiding the investigation. It presents the central concept of integrating artificial intelligence (AI) with reference class forecasting (RCF) to improve decision accuracy in Department of Defense (DoD) acquisition planning. Specifically, the study explores a hybrid AI approach that combines machine learning algorithms and natural language processing (NLP) techniques to automate reference class selection and generate probabilistic cost and schedule forecasts. The chapter also explains the significance of this approach in addressing persistent forecasting challenges such as budget overruns and schedule delays documented in Government Accountability Office (GAO) assessments. To evaluate feasibility, the research employs a qualitative design supported by quantitative techniques, including Monte Carlo simulation and forecast validation metrics such as mean absolute error (MAE) and root mean square error (RMSE). Together, these elements establish the foundation for assessing how AI-enhanced RCF may serve as a strategic tool for acquisition reform and risk-informed decision-making.

A. PURPOSE AND SCOPE

This section outlines why persistent forecasting failures within the DoD necessitate an approach that combines AI with reference class forecasting to improve early decision accuracy. Despite decades of reform, the DoD acquisition system continues to face challenges in delivering programs on time and within budget, creating risks for readiness and fiscal discipline according to GAO (2025) and Wong et al. (2022). High-profile cases such as the F-35 (GAO, 2024b) and Littoral Combat Ship (LCS) (GAO, 2022) highlight persistent forecasting weaknesses that compromise strategic outcomes.

Traditional forecasting methods often rely on analyst discretion and limited historical analogs, leaving decision-makers vulnerable to biases such as optimism bias, anchoring, and the planning fallacy, which systematically undermine forecast accuracy

(Kahneman & Tversky, 1979; GAO, 2020). These cases illustrate a broader pattern visible across major defense programs, where forecasting methods grounded in subjective judgment repeatedly fail to anticipate cost and schedule risk.

To address these persistent forecasting weaknesses in defense acquisition, this research proposes integrating AI with RCF as a potential solution. Flyvbjerg et al. (2016), states that RCF has been applied in multiple sectors, where studies indicate it may reduce optimism bias and improve budget realism, particularly in infrastructure investments. Park (2021) discusses lessons from international programs—such as the United Kingdom's (UK) mandate to apply RCF to major projects—and shows that budget overruns dropped from 38% to 5% and similar applications in infrastructure and energy have cut variance by 20–30%. This connection is relevant for the DoD because it faces similar challenges with biases, subjective analog selection, and inconsistent baseline realism. Further suggesting that the mechanisms that improved UK outcomes may also enhance early-phase forecast accuracy in defense acquisition.

Despite its recognized effectiveness, RCF has not been widely adopted within the DoD, Flyvbjerg et al. (2016) reports that this is primarily due to the difficulty of identifying statistically relevant historical analogs from vast, unstructured datasets. To address this issue, this research proposes that current AI technologies may help overcome this foundational barrier by automating the manual and time-consuming process of reference class selection, thereby making RCF scalable for the first time within the DoD. Unlike conventional RCF applications limited to budget and timeline, this study investigates whether AI-enhanced RCF has the potential to serve as a strategic tool for technology maturity forecasting, contract structuring, and life cycle sustainment forecasting.

First, DoD (2020) and GAO (2020) emphasize that technology readiness assessments (TRAs) evaluate the maturity of critical technologies elements (CTE) using structured frameworks such as Technology Readiness Levels (TRLs), with immature technologies historically linked to increased cost and schedule risk. While these assessments focus on evaluating current maturity, RCF offers a complementary approach

by incorporating empirical outcomes from analogous programs to inform probabilistic expectations of technology development.

Second, existing DoD acquisition guidance highlights that contract type selection is influenced by program risk, technological uncertainty, and cost stability considerations; however, the incorporation of empirically derived forecasting methods such as RCF into these decisions remains limited.

Third, emerging research suggests that RCF principles may be extended beyond cost and schedule estimation to support life cycle sustainment forecasting by leveraging analog system data to anticipate long-term support requirements and readiness risks. Collectively, these applications illustrate the potential for RCF to expand beyond traditional forecasting roles to inform broader acquisition decision-making and risk management processes.

The scope of this research includes evaluating current DoD forecasting practices, identifying institutional and technical barriers to AI-RCF integration, and proposing a scalable implementation framework aligned with existing policy and governance structures such as the DoD Instruction 5000 series and GAO Cost Estimating Guide. Within this scope, the following research questions are designed to examine how AI-enhanced RCF can overcome current limitations and unlock new strategic capabilities within DoD acquisition.

B. RESEARCH QUESTIONS

This study is guided by the following research questions:

1. How can AI enhance the scalability and accuracy of RCF in DoD acquisition planning?
2. What institutional, technical, and cultural barriers must be addressed to implement AI-RCF within DoD acquisition frameworks?
3. In what ways can AI-RCF be strategically applied beyond cost and schedule forecasting, such as in TRA validation, life cycle sustainment forecasting, and contract strategy optimization?
4. What measurable improvements in forecasting realism, risk management, and decision-making accountability can be expected from AI-RCF integration?

C. SIGNIFICANCE

Realistic forecasting plays a central role in maintaining readiness and fiscal discipline within defense acquisition. According to the GAO (2025), the DoD continues to experience persistent cost growth and schedule delays across its major defense acquisition portfolio. These trends are not isolated incidents but reflect systemic challenges in estimating and program execution. In particular, the report identifies a $49.3 billion increase across 30 major defense acquisition programs in a single review cycle, illustrating the scale and persistence of forecasting inaccuracies. These findings underscore the importance of improving early-phase decision quality and strengthening the analytical rigor of cost and schedule estimates.

The significance of this research is further reinforced by broader assessments of acquisition reform outcomes. The RAND Corporation (Wong et al., 2022) conducted a comprehensive review of acquisition reform efforts over a 35-year period and found that many initiatives failed to achieve sustained improvements in performance. The authors attribute these shortcomings to a lack of data-driven evaluation and insufficient tailoring of reform efforts to program-specific conditions. These findings suggest that improving forecasting accuracy requires not only methodological improvements but also the integration of empirically grounded approaches that can adapt to varying program characteristics.

Integrating AI with RCF offers a potential pathway to address these challenges. GAO (2025) emphasizes the importance of leveraging historical data and probabilistic methods to improve estimate realism, noting that traditional deterministic approaches frequently underestimate risk. Building on this insight, AI techniques—such as NLP—can be used to extract structured data from unstructured acquisition artifacts, including Selected Acquisition Reports (SARs), Modernized Selected Acquisition Reports (MSARs), and Defense Acquisition Executive Summary (DAES) reports. As GAO (2025) indicates, these data sources contain critical indicators of program performance and risk that are not consistently captured in current estimating practices.

In addition to data extraction, machine learning methods can cluster programs based on shared technical and programmatic characteristics, enabling the formation of

statistically coherent reference classes. GAO (2025) highlights that the lack of systematic use of historical analogs contributes to persistent estimation errors; therefore, automating reference class construction directly addresses a recognized gap in current practice. By combining AI-enabled data processing with RCF's empirical foundation, this approach transitions risk assessment from qualitative judgment to probabilistic analysis grounded in historical evidence. GAO (2025) suggests that incorporating probabilistic methods can improve forecasting realism, with observed improvements in accuracy ranging between approximately 20 and 30% when compared to traditional approaches.

This integrated method strengthens the traceability and defensibility of acquisition decisions. According to GAO (2025), credible estimates must be transparent, well-documented, and supported by empirical data, requirements that AI–RCF is designed to meet. By operationalizing these principles, AI–RCF supports the DoD's broader transition toward data-driven acquisition reform, consistent with guidance outlined in DoD Instruction 5000.85 and the GAO Cost Estimating Guide. Furthermore, as GAO (2025) notes, continuous incorporation of new program data enables iterative learning, allowing future forecasts to improve over time. This capability is essential for reducing systemic risk and enhancing decision-making accountability across the acquisition life cycle.

While early cost and schedule inaccuracies can cascade into long-term sustainment challenges, which according to DoD (2021) often account for 70% of total life cycle costs, this study focuses primarily on improving early-phase realism. Sustainment is addressed as crucial context and an area for future research.

Together, these elements establish the foundation for the analysis that follows. The remaining chapters examine the empirical foundations of RCF, the limitations of current forecasting practices, the proposed AI–RCF model architecture, comparative case studies, and policy recommendations for implementation.

THIS PAGE INTENTIONALLY LEFT BLANK

II. BACKGROUND AND CONTEXT

This chapter establishes the institutional and methodological context for integrating AI–RCF into defense acquisition. It begins by outlining the foundational governance structures that shape DoD forecasting practices, then examines the limitations of traditional methods and their impact on cost and schedule outcomes. The chapter also introduces key concepts that underpin the proposed integration framework developed in later chapters.

A. OVERVIEW OF DEPARTMENT ACQUISITION PROTOCOLS

The DoD acquisition system operates within a complex governance structure that organizes how capability delivery, cost control, and risk management are coordinated. DoD acquisition guidance system is designed to balance oversight, responsiveness, and accountability across the program life cycle (DoD 2020). This governance structure integrates multiple decision authorities, processes, and review mechanisms intended to ensure alignment between strategic objectives and program execution (DoD, 2015).

Three foundational processes underpin this system. The Joint Capabilities Integration and Development System (JCIDS) establish capability requirements and aligns them with strategic priorities, serving as the primary requirements generation process (DoD, 2020). The Planning, Programming, Budgeting, and Execution (PPBE) process allocates resources across the defense enterprise, linking strategic guidance to funding decisions and program prioritization (DoD, 2013). The Adaptive Acquisition Framework (AAF) provides multiple acquisition pathways that support flexible program execution, emphasizing tailoring and iterative development to accelerate capability delivery (DoD, 2020).

Although these frameworks were intended to improve agility and strategic alignment, they have not fully resolved the long-standing challenge of inaccurate forecasting. DoD acquisition policy emphasizes process compliance and milestone-based oversight as core features of the acquisition system, prioritizing structured reviews and documentation (DoD, 2020). Similarly, DoD (2013) highlights the importance of governance controls and decision authorities in ensuring accountability throughout

program execution. However, neither DoD (2015) nor DoD (2020) establishes rigorous mechanisms for enforcing analytical accuracy in cost and schedule estimates. As a result, the governance system reinforces process adherence—moving programs through required reviews—rather than improving the empirical quality of forecasting inputs.

This gap between procedural governance and evidence-based estimation represents a central weakness in the current acquisition system. While oversight structures ensure compliance and traceability, they do not systematically incorporate historical data or probabilistic methods needed to improve forecast accuracy. Consequently, forecasting errors persist despite the presence of robust governance frameworks, highlighting the need for analytically grounded approaches to estimation.

Wong et al. (2022) reinforces this assessment, finding that acquisition policy changes have consistently failed because they prioritize compliance over measurable outcomes. The authors further argue that many reform efforts are neither tailored to specific program contexts nor rigorously evaluated using empirical data, resulting in a recurring cycle of inefficiency. This structural limitation allows programs to remain fully compliant with policy requirements while still producing optimistic, deterministic estimates that underestimate risk. As Wong et al. (2022) notes, reliance on single-point estimates reinforces optimism bias and the planning fallacy, contributing directly to systemic cost growth and schedule overruns.

Schmidt (2018) provided additional insight into how these structural dynamics manifest in practice through testimony before the House Armed Services Committee (HASC). Dr. Schmidt testified that innovation within the Department of Defense often "fight [s] against entrenched processes and regulations" that have remained largely unchanged for decades (p. 1). He further noted that decision-making authority is frequently diffused, making it difficult for senior leaders to establish clear ownership of critical program decisions. According to Schmidt testimony, this diffusion reinforces slow decision cycles and institutional risk aversion, creating what can be described as an organizational inertia that inhibits the adoption of new methods, tools, and data-driven practices. His observations aligned with GAO findings that programs can satisfy all procedural requirements yet still produce baselines that fail to reflect historical

performance or underlying uncertainty, underscoring the need for forecasting approaches grounded in empirical evidence rather than procedural compliance.

GAO (2025) supports this view with portfolio reviews that show persistent cost growth and timeline slippages across major defense acquisition programs (MDAPs), with overruns exceeding $50 billion and timelines stretching to 12 years for initial operational capability (IOC). These trends underscore systemic forecasting limitations that reforms have not resolved. Portfolio-level analysis further shows that extended development timelines contribute to downstream sustainment burdens and accelerate technology obsolescence (GAO, 2025). Program-specific evidence reinforces this trend. For example, the CVN-78 carrier experienced significant schedule delays due to immature technology in its electromagnetic aircraft launch system (EMALS). In parallel, both the F-35 Joint Strike Fighter (JSF) and CVN-78 Gerald R. Ford–Class Carrier programs exceeded their original budgets by billions of dollars (GAO, 2016, 2024b; Reeves, 2025).

Recent assessments highlight a persistent gap between baseline estimates and actual outcomes for MDAPs, as budget growth has ranged from 30% to over 100% and schedule delays have extended up to 10 years (GAO, 2025). For example, GAO analysis of four MDAPs revealed cumulative overruns exceeding $50 billion and average delays of 5–7 years (GAO, 2025). Simulated AI-RCF forecasts incorporate historical variance and uncertainty bands that reveal risk patterns not visible in deterministic baseline. These forecasts demonstrate that probabilistic methods could have indicated these risks early, predicting budget growth within median (P50) and conservative (P80) confidence intervals and identifying multi-year delays before Milestone B.

These program-level examples highlight systemic weaknesses in current cost-estimating practices. To illustrate the depth of these forecasting failures, Table 1 presents a comparative analysis for four flagship programs, contrasting their initial baseline estimates against both realized outcomes and a simulated AI-RCF forecast. The results reveal a recurring and significant gap between traditional, deterministic estimates and actual performance. In contrast, the AI-RCF forecast demonstrates a much closer alignment with historical outcomes, highlighting the value of a data-driven approach that quantifies uncertainty and error explicitly.

Table 1 Comparative Outcomes of Flagship Programs.

Program	Initial Baseline (Cost/Schedule)	Actual Outcome (Cost/ Schedule)	Key Drivers of Variance	AI-RCF Simulated Forecast (Illustrative)
F-35 Joint Strike Fighter	$233 billion baseline (development + procurement); IOC ~2012	$485 billion; >5-year delay	Concurrency, immature tech, shifting requirements	P50: +60% cost growth; P80: +80% cost growth; flagged >4-year delay
LCS	~$220 million per ship; IOC ~2010	~$478 million per ship; ~10-year delay	Unrealistic modularity assumptions; scope drift	P50: +80% unit cost; P80: +100% unit cost; flagged decade-long slippage
CVN-78 Gerald R. Ford–Class Carrier	$10.5 billion estimate; delivery ~2015	>$13.3 billion; delivery slipped to 2017+	Immature EMALS, AAG, weapons elevators	P50: +20% cost growth; P80: +30% cost growth; flagged >3-year delay
Joint Light Tactical Vehicle	~$370,000–$399,000 per vehicle; IOC ~2019	~$370,000–$399,000 per vehicle; modest schedule adjustments	Mature tech, competitive prototyping, modular design	P50: Stable cost; P80: +5% cost growth; minimal delay risk

Source: Adapted from DoD (2023g), GAO (2015, 2022, 2023, 2024a, 2024b); Reeves (2025). Simulated forecast ranges are author-generated illustrations based on reference class forecasting principles and historical variance patterns identified in the cited sources.

The data presented in Table 1 illustrate that the four MDAPs experienced markedly different cost and schedule outcomes, with the largest overruns concentrated in programs that entered development with immature technologies and high integration complexity. As shown in Table 1, the JSF and CVN-78 programs exhibit the steepest cost and schedule growth, whereas Joint Light Tactical Vehicle (JLTV)—initiated with comparatively mature technologies—remained within a narrower range of variance. This pattern suggests a consistent relationship within the sampled programs: early technical maturity is associated with improved cost and schedule performance outcomes.

At the same time, the persistence of significant overruns in programs characterized by low initial readiness highlights a broader structural issue. Despite extensive DoD guidance intended to promote disciplined estimating practices, the forecasting methods currently used in acquisition do not consistently translate policy principles into realistic baseline estimates. Consequently, programs may satisfy procedural requirements while still producing systematically optimistic forecasts.

To understand this gap, it is useful to consider the Department's two principal costing references published by the DoD Cost Assessment and Program Evaluation (CAPE) office. The DoD Cost Estimating Guide v2.0 (DoD CAPE, 2022) codifies foundational methods for early-phase estimating—such as work breakdown structures (WBS), cost-estimating relationships (CERs), and formal uncertainty analysis—emphasizing traceability, documentation, and independent review. The Operating and Support (O&S) Cost-Estimating Guide (DoD CAPE, 2025) extends these principles into the sustainment phase by directing programs to integrate reliability, availability, maintainability, and data-governance considerations across the life cycle. Together, these guides establish the Department's formal framework for achieving cost realism from development through sustainment. Yet the effectiveness of these guides ultimately depends on how cost-estimating practices are implemented across the broader acquisition enterprise, where resource constraints and analytical capacity vary significantly among programs. These disparities underscore why the Department's formal cost-estimating guidance—particularly the Cost Estimating Guide and O&S Cost-Estimating Guide—plays a critical role in shaping baseline realism and establishing the analytical standards that current forecasting practices often fail to meet. This gap between established guidance and observed outcomes is evident in portfolio-level assessments of acquisition performance.

Portfolio reviews continue to document systemic expenditure growth and schedule slippage—even where programs comply with these processes on paper—indicating that deterministic baselines and inside-view judgments still dominate over empirically grounded, probabilistic forecasting (GAO, 2025). This challenge is compounded by the scale of the cost-estimating enterprise: GAO (2025) reports that more than 1,500 analysts support an annual defense budget exceeding $700 billion, producing

over 4,000 estimates each year across MDAPs and smaller efforts. The GAO report also finds that smaller programs frequently lack dedicated analytical resources and, as a result, experience 18–25% higher expenditure growth than MDAPs, with sustainment overruns exceeding $3 billion annually. These disparities underscore the need for scalable methods that can deliver MDAP-grade realism to resource-constrained efforts.

AI-RCF directly addresses this need by automating the identification of historical analogs and generating transparent, auditable probabilistic forecasts (e.g., P50/P80). In doing so, it reinforces CAPE's emphasis on documentation, traceability, and risk analysis while expanding access to rigorous forecasting beyond the largest programs. The following section examines why traditional, inside-view approaches remain vulnerable to bias and how these limitations propagate into sustainment outcomes, further motivating the case for an AI-RCF approach.

B. LIMITS OF TRADITIONAL FORECASTING AND THE SUSTAINMENT CONSEQUENCES

Traditional, inside-view forecasting approaches remain vulnerable to methodological limitations and cognitive biases, and these weaknesses propagate directly into sustainment outcomes—making them a central motivation for adopting an AI-RCF approach. Optimism bias, anchoring, and a narrow "inside view" are well-documented in cost-estimating literature and are particularly acute in analogy, parametric, and bottom-up engineering methods (Flyvbjerg, 2008; GAO, 2016). When these approaches are not rigorously benchmarked against empirical data, they produce unrealistic baselines that distort early expectations and create compounding effects throughout the life cycle, including cost growth, deferred maintenance, and readiness shortfalls (GAO, 2025).

The Cost Estimating Guide (DoD CAPE, 2022) emphasizes that credible estimates must incorporate risk and uncertainty analysis. However, recent assessments indicate continued reliance on deterministic point estimates, which aligns with observed cost growth for major defense acquisition programs. Combined total cost estimates increased by $49.3 billion (8.3%) in the past year for a subset of MDAPs assessed in consecutive reporting periods (GAO, 2025). This increase was driven largely by programs that later breached statutory cost thresholds. CAPE guidance similarly

highlights that analogy-based and parametric models are dependent on the quality and relevance of underlying data, making them susceptible to bias when applied to immature designs or novel technologies without robust historical precedent (DoD CAPE, 2022). These vulnerabilities help explain why early estimation errors routinely cascade into sustainment challenges, where inaccurate assumptions about reliability, maintenance demand, and operational tempo drive long-term affordability pressures.

Dr. Schmidt's testimony further highlights how upstream requirements processes amplify these forecasting weaknesses. He argued that the Department's requirements system—originally designed for long-cycle hardware development—has become "the single greatest barrier to rapid technological advancement" (Schmidt, 2018, p. 2). This rigidity delays early decision-making, constrains the ability to incorporate emerging data, and reinforces deterministic planning assumptions that fail to reflect technical maturity or historical variance. Schmidt's critique aligns with GAO findings that requirements instability and premature commitment to immature technologies are major contributors to cost growth and schedule slippage across MDAP portfolios. Together, these insights underscore how structural features of the requirements process distort early baselines and create conditions where traditional forecasting methods systematically underestimate risk.

The limitations of traditional forecasting methods extend well beyond development phases, driving significant life cycle cost growth and readiness degradation during sustainment. This connection emerges because early-phase estimating assumptions shape downstream maintenance and support requirements. Under-scoped planning and reliance on narrow analogs systematically underestimate long-term support requirements. Sustainment challenges further illustrate how early forecasting weaknesses compound over the life cycle. GAO (2023) assessments report that the F-35 fleet's mission-capable rate remained near 55% in 2023, with roughly 10,000 components awaiting depot-level repair—evidence of persistent bottlenecks in the sustainment enterprise. These pressures are magnified by the program's long-term cost profile: of the F-35's estimated $1.7 trillion life cycle cost, approximately $1.3 trillion is attributed to O&S, underscoring how sustainment dominates total ownership costs. Portfolio-level analysis also shows that MDAPs require an average of 12 years to reach IOC, extending

exposure to sustainment risk and increasing the likelihood that early estimation errors will cascade into affordability challenges (GAO, 2024a). Together, these outcomes demonstrate how unrealistic baselines established during development can drive long-term readiness shortfalls and cost growth, reinforcing the need for forecasting methods that incorporate empirical risk evidence rather than deterministic assumptions.

Collectively, the GAO sustainment findings, cost-growth trends, and readiness indicators illustrate how fragmented governance, underdeveloped cost baselines, and unresolved technical risks propagate from acquisition into sustainment when forecasting methods fail to incorporate robust, out-of-sample evidence. Deterministic point estimates create an illusion of precision that masks underlying uncertainty, contributing to the cascading cost growth and schedule delays seen across the portfolio. In contrast, a probabilistic approach reveals the full spectrum of potential outcomes, enabling decision-makers to quantify and manage risk. Figure 1 visually contrasts these two methodologies by comparing a traditional single-point estimate—represented as a fixed cost value that obscures uncertainty—with a probabilistic forecast illustrated as a distribution of possible outcomes, including best-case, expected, and worst-case scenarios. While single-point forecasting presents a narrow and deterministic view of program cost, the probabilistic approach highlights the range of potential life cycle costs and enables more informed, risk-aware decision-making.

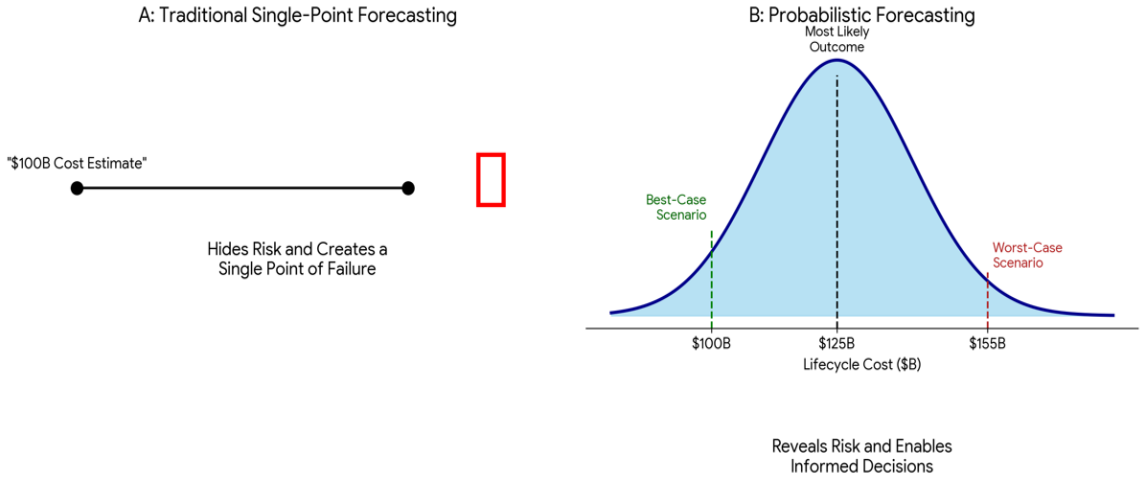

Figure 1. Traditional Versus Probabilistic Forecasting Approaches. Adapted from DoD (2023g) and GAO (2015, 2020, 2025). Image generated using Microsoft Copilot (2026).

AI-RCF offers a scalable solution by using historical analogs and probabilistic modeling to forecast both development outcomes and long-term sustainment liabilities. By generating probabilistic estimates for life cycle costs and maintenance demand, AI-RCF enables decision-makers to anticipate sustainment risks earlier, allocate resources more effectively, and improve life cycle budget realism across the acquisition-to-sustainment continuum.

To appreciate the consequences of relying on traditional estimating methods, it is useful to examine how these approaches have performed in practice. Table 2 provides this perspective by comparing optimistic baselines against actual outcomes for flagship programs, highlighting how inadequate risk treatment drives cascading effects—massive cost growth, multi-year schedule delays, and sustainment shortfalls.

Table 2 Consequences of Inside-View Estimating Across Flagship Programs

Program	Original Baseline	Current Acquisition Cost	Life cycle Cost	Schedule Impact	Key Sustainment Issues
F-35 Joint Strike Fighter	$233 billion	~$485 billion	~$2.1 trillion	Block 4 ≥5 years late; 110 aircraft delivered in 2024 averaged 238 days late	>10,000 components awaiting depot repair; mission-capable rate ~55%
Ford-Class Carrier (CVN-78)	$10.5 billion (lead ship)	$12.9 billion	N/A	Lead ship slipped >2 years; CVN-79 now $13.2 billion with added cost-to-complete	Persistent technology, immaturity, and concurrency risks
LCS	$25 billion (planned through FY2035)	$31 billion (procurement)	$60 billion O&S	Multiple sea frames delivered late; mission packages years behind	Heavy contractor reliance; incomplete sustainment plan; readiness gaps

JLTV	$6.7 billion (LRIP)	$8.66 billion (follow-on)	N/A	Transition to AM General caused 6-month slip; full-rate production now mid-2026	Parts shortages, depot overhaul reductions threaten mission-capable rates

Source: Adapted from GAO program assessments of F-35, Ford-class aircraft carrier, LCS, and JLTV programs (GAO, 2015, 2022, 2023, 2024b, 2025). Specific cost, schedule, and sustainment data are drawn from the respective system-level reports.

Across these programs, Table 2 demonstrates a consistent pattern of cost growth, schedule slippage, and sustainment challenges that extend beyond individual program characteristics. As reported by GAO (2024b), the F-35 baseline nearly doubled in cost, while the Ford-class carrier experienced substantial cost increases and multi-year schedule delays (GAO, 2015). Similarly, the LCS and JLTV programs exhibited persistent performance issues and sustainment shortfalls, although to varying degrees (GAO, 2022, 2023).

These results indicate that programs relying on optimistic initial assumptions and narrow risk characterization fail to account for the full distribution of historical outcomes. As a result, early baseline estimates systematically underestimate cost and schedule risk. This pattern highlights a broader structural limitation within current estimating practices, wherein programs may adhere to established processes while still producing forecasts that do not reflect empirical program performance.

Although analogy, parametric, and engineering approaches remain core GAO and CAPE best practices, GAO (2020) emphasizes that credible estimates must be comprehensive, well-documented, accurate, and independent, with explicit risk and uncertainty analysis—requirements that are often unmet when programs depend heavily on judgment or limited historical analogs (DoD CAPE, 2022). As the F-35 and Ford-class cases demonstrate, insufficient risk treatment leads to inflated confidence in early baselines and subsequent unraveling as technical, integration, and sustainment realities emerge (GAO, 2015, 2024b, 2025).

Traditional forecasting methods differ in technique, but they share a structural limitation: they rely on program-specific assumptions that systematically understate

uncertainty (GAO, 2025). Analogy-based estimating illustrates this weakness most clearly. Although analogy appears similar to reference-class reasoning, it typically selects one or a small set of "comparable" programs—a subjective process that anchors the estimates to favorable precedents and obscures the full distribution of historical outcomes according to Flyvbjerg (2008). DoD CAPE (2022), demonstrates that this narrow framing is not unique to analogy; parametric and engineering-based approaches exhibit similar vulnerabilities when applied without robust uncertainty analysis. GAO (2025) indicates that these limitations help explain why early baselines often diverge sharply from realized performance.

According to Flyvbjerg (2021), RCF addresses this structural weakness by replacing individually chosen analogs with the entire empirical distribution of outcomes from a statistically coherent reference class. Instead of relying on subjective comparator choice, RCF draws on evidence and probabilistic ranges that capture variance traditional methods routinely miss. AI-RCF strengthens this foundation further by automating reference-class construction and generating transparent, auditable risk-adjusted forecasts. Together, these advantages demonstrate why analogy-based estimation resembles RCF in concept but lacks the empirical rigor and predictive reliability that RCF—and especially AI-RCF—provides.

C. REFERENCE CLASS FORECASTING

Given these limitations in traditional forecasting methods, RCF provides a disciplined way to counter the estimation errors documented in earlier sections by anchoring forecasts in empirical outcome distributions rather than program-specific assumptions. Its relevance for defense acquisition lies in how it reframes early estimates: Flyvbjerg et al. (2016) reports that instead of relying on optimistic technical baselines, RCF requires programs to benchmark against the actual performance of comparable efforts, revealing risk patterns that traditional methods often overlook. GAO (2020) finds this outside-view discipline has produced measurable improvements in budget and schedule realism across large, complex projects, particularly where technical uncertainty and integration challenges resemble those found in major defense programs.

For the DoD, RCF's primary value is its ability to translate historical performance into actionable probability bands—such as median (P50) and conservative (P80) estimates—that better capture the true range of potential outcomes. These empirically grounded distributions strengthen milestone decisions, contingency planning, and early trade-space analysis. As mentioned earlier, the evidence from the United Kingdom reinforces this point: Park (2021) found that after RCF became mandatory in 2003, budget overruns on major projects declined sharply and forecasting accuracy improved across government portfolios. As subsequent sections show, integrating AI with RCF further enhances these benefits by automating reference-class construction and scaling the method across diverse acquisition portfolios.

Building on these results, two core strengths explain why RCF has become a preferred approach in complex project environments:

- Bias Mitigation: Shifts focus from internal estimates to external benchmarks.
- Transparency: Provides a defensible, data-driven basis for forecasts.

These principles also appear relevant to defense acquisition, including both MDAPs and selected smaller rapid-prototyping efforts, particularly where early estimates rely heavily on subjective judgment. Although RCF offers a structured, empirically grounded alternative to traditional forecasting, its practical implementation within the Department remains constrained by what Defense Acquisition University (DAU) identifies as the reference class problem—the difficulty of identifying statistically relevant analogs within large, heterogeneous historical datasets (DAU, 2024). DoD CAPE (2025) finds this problem stems directly from the fragmentation of acquisition data across disparate systems like Cost Assessment Data Enterprise (CADE), Enterprise Visibility and Management of Operating and Support Costs (EVAMOSC), and MSARs. GAO (2025) finds that assembling even a basic dataset from these sources requires extensive manual effort, making the process labor-intensive and nearly impossible to scale, leaving the method's full potential untapped.

This is precisely the implementation gap that AI is poised to fill. AI enables RCF to be applied at scale by identifying patterns across historical programs, clustering them based on technical and programmatic similarity, and generating probabilistic P50/P80

forecasts that reflect the full distribution of historical outcomes. By automating these processes, AI enhances the precision and consistency of RCF and embeds empirical rigor into acquisition decisions. This sets the stage for examining how AI can expand the applicability of RCF across diverse acquisition portfolios.

D. ARTIFICIAL INTELLIGENCE IN FORECASTING

Because RCF depends on identifying statistically relevant historical analogs, its effectiveness is constrained by the manual effort required to construct reference classes. AI techniques, like machine learning (ML), NLP, and predictive analytics, help overcome this limitation by grouping programs into statistically coherent clusters and pattern extraction. This capability addresses a critical gap in defense acquisition, where manual reference class construction is labor-intensive and prone to bias.

The scale and complexity of defense acquisition data create significant barriers to timely, high-quality analysis. A single MSAR can exceed 1,000 pages, and DAES reports often contain hundreds of qualitative risk statements requiring manual interpretation (DoD CAPE, 2025). GAO (2024a) audits similarly find that cost analysts spend up to 70% of their time on data collection and normalization rather than analysis, reducing the time available for risk mitigation and delaying milestone decisions. These challenges are even more acute in smaller programs, which often lack dedicated analytical resources and operate under compressed timelines.

AI techniques offer a path to alleviating these constraints by automating data ingestion, extracting structured features from unstructured acquisition documents, and enabling more consistent application of probabilistic methods. Recent research demonstrates that artificial intelligence is already being applied in high-stakes domains, including logistics optimization, predictive maintenance, and cybersecurity risk assessment (Khan et al., 2024; Shamim et al., 2025; Su et al., 2024). For example, Welch (2025) reports that platforms such as Palantir's Foundry and Gotham—recently consolidated under a $10 billion Army contract—illustrate how AI-driven analytics can support pattern recognition and operational planning at enterprise scale. Integrating AI with RCF has been proposed as a way to automate reference class formation, enhance pattern recognition, and produce dynamic probabilistic forecasts that adapt as new data

becomes available. This method aligns with the principles outlined in the DoD Cost Estimating Guides (DoD DCAPE, 2022, 2025) which emphasize three pillars of credible estimation: grounding forecasts in historical data, ensuring transparency and traceability, and incorporating risk and uncertainty analysis.

AI-enabled RCF operationalizes these requirements by automating the use of empirical analogs, generating auditable outputs that reflect real-world variability. In doing so, it not only meets CAPE's standards but extends robust forecasting capabilities to smaller programs that often lack dedicated analytical resources, closing a critical gap in life cycle risk management. These AI capabilities contribute to a more comprehensive assessment for applying probabilistic, evidence-based methods to technology maturity assessments.

E. STRENGTHENING TECHNICAL RISK ASSESSMENT WITH AI-RCF

A key contributor to program risk is the inconsistent and often subjective execution of TRAs, which can lead to critical underestimation of technical maturity and integration risk. The consequences are significant and well-documented. According to GAO (2020), programs that proceed with technologies below appropriate maturity thresholds face substantially higher risks of cost growth and schedule slippage, as immature technologies require unforeseen design iterations, integration work, and testing cycles.

This direct link between technological immaturity and poor outcomes is evident in flagship programs. For example, the Ford-class aircraft carrier and the F-35 JSF both entered development with multiple CTEs at TRL 4–5, resulting in cumulative cost overruns exceeding $13 billion and schedule delays of more than five years (GAO, 2015. 2024b, 2025; Reeves, 2025).

Figure 2 shows that programs entering development with lower Technology Readiness Levels (TRL 4–5) tend to experience significantly higher cost growth than programs with more mature technologies (TRL 6–7+). Programs below TRL 6, in particular, are associated with substantially higher expenditure growth, often exceeding 30%. This pattern underscores a key limitation of TRLs as a forecasting tool. Because

TRLs capture only a single dimension of technical maturity, they do not fully account for broader integration risks or system-level dependencies. As a result, programs may appear technically mature while still carrying substantial downstream cost and schedule risk.

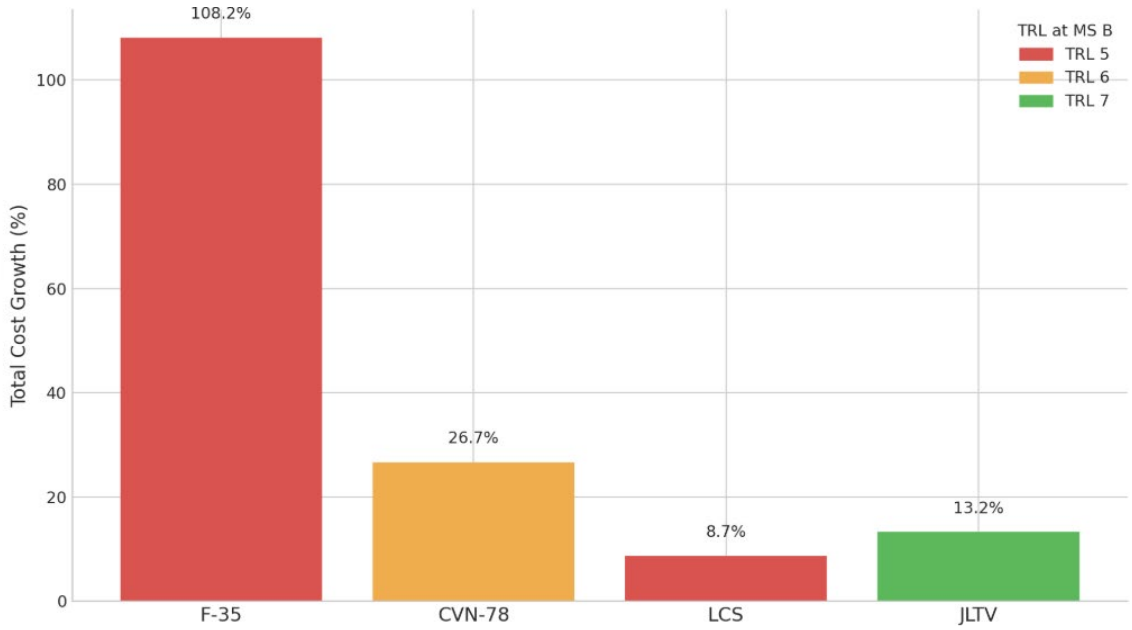

Figure 2. Technology Readiness Level at Milestone B versus Cost Growth. Adapted from GAO (2015, 2020, 2025) and DoD MSARs (2023-2024). Image generated using Microsoft Copilot (2026).

This analytical gap contributed to the statutory requirement for independent technical risk assessments (ITRAs) to provide a more comprehensive evaluation (Independent Technical Risk Assessments for Major Defense Acquisition Programs, 2023). Yet, in practice, both TRAs and ITRAs often rely on subjective judgment rather than data-driven analysis. GAO (2020, 2025) reviews indicate that ITRAs frequently lack analytical rigor, with over 60% of those reviewed between 2018 and 2023 citing qualitative assessments as the primary basis for risk ratings and fewer than 15% incorporating probabilistic modeling.

This is precisely the shortfall that AI-RCF is designed to address. Instead of relying on subjective judgment, AI-RCF grounds readiness assessments in empirical data. By analyzing historical outcomes from programs with comparable complexity and integration challenges, it can generate risk-adjusted probabilistic forecasts. For example, rather than a binary "ready/not ready" judgment, the assessment could state: "There is an

80% probability that this radar system will not reach TRL 7 before the Engineering and Manufacturing Development phase." This level of granularity enables milestone decision authorities to tailor acquisition strategies, allocate contingency budgets, and make trade-offs based on quantified uncertainty rather than subjective optimism, directly fulfilling the analytical rigor envisioned by both GAO (2020) and DoD CAPE (2022).

In summary, the DoD faces recurring forecasting challenges rooted in methodological limitations, cognitive biases, and a lack of TRA rigor. Having established how AI-RCF can address these systemic issues, the following literature review will examine the empirical foundations of RCF and explore how AI can overcome its traditional scalability barriers, providing the analytical grounding necessary to evaluate its feasibility within defense acquisition.

III. LITERATURE REVIEW

Building on the reference class problem identified earlier, this literature review situates the present study within the broader scholarship on forecasting methodologies, acquisition reform, and AI. To do so, the review first examines the empirical foundations of RCF, from its theoretical origins to evidence from cross-sector adoption and key scholarly debates on its limitations. It then evaluates how these insights establish the analytical groundwork for an AI-enabled approach tailored to the unique data and decision-making environments of the DoD.

A. EMPIRICAL FOUNDATIONS OF REFERENCE CLASS FORECASTING

This section reviews the empirical foundations of RCF, beginning with its theoretical origins and validation across sectors, before examining evidence from policy-level adoption.

1. Origins and Empirical Validation of RCF

The empirical foundations of RCF are rooted in behavioral economics, drawing on Kahneman and Tversky's (1979) "outside view" concept and the structured methodology later operationalized by Flyvbjerg et al. (2016). Subsequent scholarly work provides evidence that anchoring forecasts in empirical distributions can help counter optimism bias and the planning fallacy inherent in project-specific assumptions.

Flyvbjerg et al. (2016) analyzed 2,000 global megaprojects and found that 90% exceeded initial budget estimates, with average overruns of 28% for transport, 45% for energy, and over 100% for information technology systems. Projects applying RCF reduced variance by up to 30% compared to traditional forecasting methods (Internal Consulting Group [ICG], 2016). These findings provide a substantial empirical basis for RCF as a bias-mitigation tool and set the stage for examining policy-level adoption.

2. Policy Adoption and Defense Relevance

A deeper examination of the UK case provides compelling quantitative evidence of RCF's impact. After mandating RCF for major infrastructure investments in 2003,

Park (2021) finds that average cost overruns fell from 38% to 5% for total procurement and from 47% to 4% for traditional procurement, representing marked improvements in budgeting accuracy. Park also shows that the UK exceeded its budget-accuracy target by 12%, whereas the United States—which has not adopted RCF—underperformed by 17%, underscoring the role of governance in converting forecasting discipline into measurable performance gains.

Evidence beyond the UK reinforces this pattern. Transportation and energy portfolios using probabilistic methods have achieved 15–25% improvements in budget adherence and 10–18% reductions in delivery delays, while RCF-based approaches in information technology programs reduced schedule slippage by up to 20% (Park, 2021; Flyvbjerg et al., 2016; ICG, 2016). Together, these cross-sector results demonstrate that RCF's benefits are not domain-specific but arise from a consistent shift toward empirical, distribution-based forecasting.

These findings suggest that RCF can serve as an effective countermeasure to both cognitive and institutional sources of bias, particularly when supported by strong data governance practices. Cantarelli et al. (2025) reports that large-scale analyses of infrastructure portfolios demonstrate that projects applying RCF techniques achieve up to 40% greater cost predictability and 25% higher schedule adherence compared to those using conventional deterministic methods. Similarly, Park (2021) finds that RCF-based frameworks improve budget control and reduce performance shortfalls across transportation and energy programs.

Taken together, these findings reinforce the argument that probabilistic, data-driven forecasting methods improve acquisition outcomes when systematically applied, reinforcing the relevance of integrating RCF into early-phase cost and schedule planning.

Although Park's analysis focuses on civilian infrastructure, the underlying mechanisms that drive forecasting improvements—optimism bias, strategic misrepresentation, and baseline inaccuracy—parallel those documented across U.S. defense acquisition portfolios (GAO, 2025; Wong et al., 2022). This alignment highlights the relevance of RCF for DoD programs, where persistent cost growth and schedule slippage reflect many of the same behavioral and methodological shortcomings.

The potential fiscal implications are substantial. With a major project portfolio exceeding $1 trillion annually (GAO, 2024a), applying the 17% underperformance gap identified by Park (2021) suggests that systematic adoption of RCF has the potential to yield approximately $170 billion in annual savings. Taken together, these findings illustrate the strategic importance of embedding empirical forecasting methods into U.S. acquisition governance to improve budget accuracy, reduce systemic cost growth, and strengthen decision accountability.

B. THEORETICAL DEBATES AND METHODOLOGICAL REFINEMENTS

While the empirical evidence highlights RCF's practical effectiveness, the methodology is also the subject of ongoing scholarly debate and refinement. This section examines critiques regarding its application, explores methodological advancements designed to enhance its rigor, and discusses the challenges that have limited its adoption, particularly within the DoD.

While these results underscore RCF's practical effectiveness, they also raise deeper questions about how success should be measured. A newer stream of scholarship cautions against equating forecast error with welfare loss and urges careful benefit-cost interpretation of the "iron law" literature—Flyvbjerg's empirical finding that major projects are consistently over budget, over time, and underperforming on benefits (Välilä, 2024). This tension—between evidence supporting bias mitigation and debates about welfare conclusions—helps frame the theoretical relevance of RCF for defense acquisition. Understanding this debate is important because adoption decisions in the DoD hinge not only on technical validity but also on governance and welfare considerations. Policy-level adoption evidence, such as UK practice, indicates that governance changes can convert RCF benefits into measurable performance gains, highlighting the importance of institutional context in realizing its full potential.

A recent peer-reviewed synthesis by Cantarelli et al. (2025) consolidates what is known about RCF across sectors: it reliably counters optimism bias when implemented with discipline, but its performance is not uniform—it hinges on reference-class validity (how classes are defined and features selected), percentile choice (uplift selection), and data quality. Critically, the authors call for domain-specific back testing—including

defense acquisition portfolios—to verify that uplift distributions and risk bands hold under the technical novelty, heterogeneous data, and governance incentives typical of MDAPs. This matters because, despite decades of reform, defense programs still exhibit persistent schedule growth and only modest improvements in cost-variance, signaling that inside-view baselines continue to underrepresent risk. In such settings, outside-view methods must be validated against defense-specific evidence, not borrowed wholesale from infrastructure casebooks (Cantarelli et al., 2025; Jones et al., 2024).

Accordingly, this capstone project responds to that call by proposing a defense-oriented application of AI-enabled RCF, emphasizing transparent validation using established performance metrics such as MAE, RMSE, and confidence interval coverage. Rather than assuming transferability from infrastructure domains, this approach is tailored to the data constraints, technical heterogeneity, and governance requirements of DoD acquisition, demonstrating how probabilistic forecast validity can be assessed prior to institutional adoption. These developments complement ongoing methodological refinements discussed in the literature, including weighting strategies to reduce uplift inflation and integration with machine learning to better capture complex risk patterns and improve predictive accuracy.

Building on these critiques, scholars have proposed refinements and hybrid approaches to enhance RCF's predictive rigor. Zani and Adey (2025) introduce weighted-RCF variants that apply differential weights to historical analogs based on relevance and data quality, producing more conservative uplift distributions and reducing the risk of "over-uplifts" that inflate budgets unnecessarily. This evolution matters in defense acquisition because excessive contingency can be as damaging as optimism bias, often triggering affordability concerns and political pushback (GAO, 2023). For MDAPs, where cost credibility is a statutory requirement, weighting mechanisms underscore the importance of rigorous reference-class construction and percentile selection. Integrating these bias-mitigation and governance principles into AI-enabled RCF helps ensure automation does not replicate outdated practices but instead supports emerging best practices for balanced and defensible forecasts.

Beyond weighting strategies, scholars have begun pairing RCF with machine learning to address distributional complexity in large-scale projects. Natarajan (2022) demonstrates that integrating ML with RCF improves the ability to capture fat-tailed budget and schedule overrun patterns—risks that conventional uplift curves often understate. Together, these developments underscore why automation should not merely replicate legacy RCF but integrate emerging best practices for uncertainty treatment and predictive modeling. These distributional findings matter for defense programs because they frequently exhibit extreme variance driven by technology immaturity, concurrency, and shifting requirements (GAO, 2023; Wong et al., 2022), making traditional percentile-based uplifts insufficient for realistic planning. By leveraging ML's capacity to model nonlinear relationships and tail risk—capabilities documented in recent defense analytics research (Wong et al., 2022)—AI-enabled RCF can move beyond static uplift tables toward dynamic, data-driven forecasts that reflect real-world uncertainty. This empirical precedent reinforces the rationale for the hybrid approach advanced in this capstone project, ensuring that automation delivers not only scalability but also improved predictive fidelity for high-stakes defense portfolios.

Despite notable successes abroad, RCF adoption remains uneven. Many international governments and private-sector organizations have institutionalized RCF as a standard for major capital projects, citing reductions in budget overruns of up to 50% in the UK after its 2003 mandate (Park, 2021). In contrast, the DoD has yet to formally integrate RCF into acquisition policy, relying instead on traditional compliance-driven frameworks. Wong et al.'s (2022) synthesis concludes that uniform, one-size-fits-all reforms rarely deliver measurable improvements; rather, strategies must be tailored to program complexity and supported by rigorous, data-driven evaluation. AI-enabled RCF offers a direct response to these gaps by grouping programs into statistically coherent clusters, leveraging historical performance distributions, and embedding empirical evidence into probabilistic forecasts. This approach aligns with call for adaptive acquisition pathways and analytic rigor, while addressing persistent optimism bias that contributes to average budget overruns of 20–45% in defense programs and schedule slippages exceeding 25% in information technology and software acquisitions (GAO, 2023; Wong et al., 2022).

In defense contexts, Jones et al. (2024) reports that portfolio-level analyses consistently show budget overruns and timeline extensions across MDAPs, reinforcing the need for probabilistic methods early in the life cycle. Jones (2022) similarly finds that variability in cost growth has narrowed over decades, but typical programs still experience slippage, implying persistent optimism and structural drivers unaffected by compliance-focused reforms alone. Analogous patterns in MDAP data suggest that RCF's empirical logic transfers to defense contexts when maturity and baseline realism are enforced.

The difficulty of consistently identifying valid analogs across heterogeneous defense programs underscores why manual RCF processes struggle to scale and why automation is required for institutional adoption. Practitioner studies, including global megaproject analyses, show that RCF's predictive value depends on rigorous reference-class construction and disciplined probabilistic modeling. However, because defense programs vary widely in technical scope, technology maturity, and operational context, defining a statistically valid reference class requires substantial manual effort. These challenges are well-documented in the ICG (2016) case studies. While these studies provided a structured approach—defining the class, selecting metrics, and generating forecasts—the underlying manual constraints highlight why automated approaches are necessary for large and diverse defense portfolios. As a result, scalable, outside-view baselines become essential for mitigating systemic forecasting bias in defense acquisition.

The literature identifies several limitations in manual RCF implementation, including scalability constraints, analog-selection challenges, and the need for more consistent probabilistic modeling. Recent studies highlight emerging applications of ML and NLP that may help automate reference-class construction and improve forecasting precision, though empirical validation in defense contexts remains limited. These strands of research collectively point to an evolving intersection between behavioral forecasting methods and AI-enabled analytics.

As previously established in Park's (2021) analyses, RCF's empirical performance in the UK demonstrates how governance can convert outside-view

forecasting into measurable results. Table 3 summarizes these impacts and contrasts them with U.S. performance, illustrating how institutional adoption shapes cost realism and budget discipline.

Table 3 Impact of Reference Class Forecasting Adoption on Cost Overruns and Budget Performance.

Country	RCF Adoption Year	Avg. Cost Overrun (Before RCF)	Avg. Cost Overrun (After RCF)	Budget Performance
UK	2003	38%	5%	+12% over target
US	Not adopted	~30% (MDAP portfolio average)	No change	−17% under target

Source: Adapted from Flyvbjerg et al. (2016); GAO (2025); Park (2021).

Table 3 highlights the contrast between RCF adoption in the United Kingdom and its absence in the United States, illustrating the measurable impact of probabilistic, outside-view forecasting on cost realism and budget performance. These results reinforce the empirical evidence presented earlier, demonstrating that structured reference class methods improve acquisition outcomes when systematically applied.

These findings are further reinforced by broader meta-analyses of infrastructure portfolios, which demonstrate that projects applying RCF achieve up to 40% greater cost predictability and 25% higher schedule adherence compared to conventional deterministic methods according to Cantarelli et al. (2025). Beyond cost control, Park (2021) states that RCF frameworks have been linked to improved benefit realization, reducing performance shortfalls by nearly 50% in transportation and energy programs. Park's (2021) research concludes that these outcomes reinforce the principle that structured reference classes and probabilistic forecasting are not domain-specific but scalable across sectors, including information technology implementations where failure rates are historically high. exceed 60%, overruns exceed 45%, and benefits shortfalls reach 56% without such interventions (ICG, 2016). This limitation underscores the need for scalable, data-driven solutions.

Emerging AI capabilities, including machine learning and natural language processing, offer a practical path forward. Anton et al. (2020) reports that these tools can ingest historical acquisition data, cluster programs by shared characteristics, and generate

probabilistic forecasts that reflect empirical outcome distributions. DoD (2025) further reports that by combining RCF's empirical rigor with AI's scalability, AI-enabled RCF represents a significant methodological advancement, enabling a shift from manual, subjective forecasting toward dynamic, data-driven analysis. This integration directly addresses the reference-class identification bottleneck, making RCF more feasible for application across complex defense acquisition portfolios.

The empirical record converges on three points: (1) the inside view structurally underestimates risk; (2) RCF improves realism when applied with discipline and high-quality classes; and (3) scholarly debate persists regarding how forecast accuracy should be interpreted in terms of broader decision outcomes and policy implications. These findings frame the relevance of RCF for defense portfolios, where technical novelty and institutional incentives amplify bias and uncertainty. The next section turns from RCF's empirical foundations to the limitations of traditional forecasting methodologies used in DoD practice.

C. FORECASTING METHODOLOGIES AND LIMITATIONS

The academic and government literature extensively documents the limitations of the traditional cost-estimating methods used across DoD practice. Peer-reviewed studies and federal audits consistently show that traditional techniques are susceptible to optimism bias, anchoring, and an overly narrow "inside view," especially in complex programs with sparse historical data (Flyvbjerg, 2008; GAO, 2024a). Reliance on deterministic point estimates also systematically understates uncertainty, contributing to the cost and schedule breaches repeatedly highlighted in GAO and CAPE guidance. Conversely, advances in probabilistic forecasting offer more rigorous treatment of uncertainty. For example, methods such as Monte Carlo simulation and quantile regression generate distributions—P50, P80, and tail-risk bands—rather than single-point baselines. Empirical research reinforces the value of probabilistic forecasting: Chronopoulos et al. (2024) show that deep neural quantile regression improves tail-risk capture relative to linear approaches, while Fitzenberger and Wilke (2015) demonstrate extensions that better handle censoring and nonlinearity. These findings are directly relevant to defense acquisition, where MDAP cost-growth data frequently display skewed

or fat-tailed patterns that deterministic methods fail to represent (GAO, 2023; Wong et al., 2022). Incorporating probabilistic methods aligns with GAO and DCAPE recommendations on uncertainty analysis and establishes an analytical foundation for both RCF and emerging AI-enabled forecasting approaches. Nevertheless, adoption remains slow due to cultural inertia, fragmented data systems, and the labor-intensive nature of manual probabilistic modeling.

RCF addresses several of these shortcomings by providing a structured outside view grounded in empirical performance data. Despite this data, studies note that RCF remains difficult to apply consistently across heterogeneous programs, largely because constructing valid reference classes and extracting comparable features from acquisition documentation is resource-intensive. Recent work demonstrates that AI techniques, particularly machine learning and natural language processing, can automate the extraction of structured features from unstructured acquisition artifacts (Su et al., 2024), enabling more scalable and adaptive forecasting workflows. These capabilities support greater consistency and broader applicability across heterogeneous program types. As a result, AI-enabled RCF emerges as a promising pathway for overcoming the scalability constraints that have historically limited its adoption within defense acquisition.

Evidence from cross-sector studies further demonstrates the value of AI-enabled forecasting. Shamim et al. (2025) report that AI-driven models improve budget estimation accuracy by 75–90% compared to deterministic baselines. Within aerospace and defense, Khan et al. (2024) document expanding use of machine learning for schedule prediction and risk assessment, though they caution that governance and explainability remain critical for institutional acceptance. Natarajan (2022) shows that combining machine learning with RCF yields more statistically consistent modeling of extreme outcomes, and Jimenez et al. (2016) find that probabilistic approaches applied to pre-Milestone B variables significantly improve schedule-duration prediction. Collectively, these studies strengthen the case for early integration of AI-enabled forecasting in defense acquisition.

A broad set of machine-learning techniques—gradient-boosting machines, random forests, and neural networks—has demonstrated strong predictive performance in

cost and schedule estimation. According to DAU (2024) gradient boosting enhances accuracy by combining weak learners, random forests mitigate overfitting through ensemble diversity, and neural networks capture nonlinear relationships often present in defense program data. Welch (2025) further states that commercial platforms such as Palantir's Foundry and Gotham illustrate how integrated data environments support large-scale predictive analytics, enabling applications ranging from logistics optimization to predictive maintenance and portfolio risk modeling. Within DoD estimating practice, however, analogy, parametric, and engineering/bottom-up approaches remain dominant despite their vulnerability to bias when historical data are sparse or technologies immature. Systems-engineering literature, including COSYSMO research, emphasizes the need for calibrated models and robust uncertainty treatment to avoid these pitfalls (Valerdi, 2010; SEBoK, 2024).

Despite progress in adjacent domains—such as cyber risk scoring, supply-chain resilience, and operational planning—AI integration into acquisition forecasting remains limited. DAU recognizes the potential for AI-enabled estimating, but formal frameworks and standardized methods are still emerging (DoD, 2025). Meanwhile, peer-reviewed acquisition research shows that schedule growth has not improved significantly over several decades, even as cost-growth variance narrows (Jones et al., 2024), underscoring the limits of deterministic methods at Milestone B. RCF directly addresses these constraints by grounding forecasts in empirical distributions, yet scholars also identify challenges related to class validity, selection subjectivity, and data quality—issues that become acute in heterogeneous defense portfolios (Cantarelli et al., 2025).

AI meaningfully mitigates many of these barriers. NLP can mine DAES reports, MSARs, and milestone documents for latent patterns, while unsupervised learning can cluster programs based on statistical similarity rather than subjective judgment. This accelerates reference-class construction, reduces manual bias, and increases consistency across acquisition portfolios according to DoD (2025). AI-enhanced RCF also aligns with the principles outlined in the 2025 DoD Cost Estimating Guide (DoD CAPE, 2025), which emphasize empirical rigor, transparency, and probabilistic analysis. For example, clustering historical aircraft programs by their cost-growth distributions can inform more realistic forecasting for new fighter platforms.

Traditional methods will remain part of the estimating toolkit, but they are no longer sufficient on their own. The methodological frontier lies in integrating CERs with outside-view evidence and rigorous uncertainty quantification. This fusion strengthens the analytical foundation for defense estimating while also introducing governance, data quality, and explainability challenges that shape practical implementation.

Building trust and ensuring credible governance over this new analytical frontier requires a human-on-the-loop approach—where AI augments rather than replaces expert judgment—is essential to ensure accountability, trustworthiness, and compliance with GAO and DoD standards. While AI offers a powerful response to the scalability and bias challenges inherent in RCF, its adoption within the DoD remains constrained by institutional, cultural, and technical barriers. These constraints underscore the need for governance frameworks that enable transparent, auditable, and trustworthy forecasting.

D. INSTITUTIONAL BARRIERS AND REFORM OPPORTUNITIES

Organizational and institutional dynamics exert a profound influence on forecasting accuracy in defense acquisition. Despite decades of reform—from the Packard Commission's 1986 call for streamlined oversight and professionalized program management to today's AAF—forecasting failures remain pervasive across the acquisition life cycle. GAO (2024a) evaluations consistently reveal deep-rooted flaws in DoD acquisition practices, citing chronic cost escalation, prolonged timelines, and weak risk controls. Wong et al.'s (2022) extensive analysis of acquisition research reinforces this pattern; their findings conclude that repeated policy cycles often fall short because they prioritize compliance over measurable outcomes and lack rigorous evaluation of effectiveness. This gap between policy design and analytic rigor exposes a structural weakness: governance frameworks define processes but often do not enforce empirical methods for estimating financial and schedule risk. Compounding this weakness, cultural incentives can encourage reliance on fixed-point estimates rather than probabilistic evidence, perpetuating optimism bias and the planning fallacy (Flyvbjerg, 2021; Prater et al., 2017).

1. Persistent Structural Drivers

Institutional barriers are not limited to technical methods; they are embedded in governance structures and cultural incentives. Peer-reviewed acquisition research reinforces this pattern. Porter et al. (2009) and McNicol's (2022) research revealed that funding instability, requirement churn, and concurrency decisions amplify program risk, while organizational incentives often reward optimistic baselines rather than realistic planning. McNicol (2022) shows that structural drivers and incentive misalignments persist across policy cycles, and Ahn and Menichini (2022) find that workforce behavior and retention dynamics further constrain the effectiveness of reform efforts. These institutional frictions help explain why, even amid repeated attempts at process improvement, empirical trend analyses show that MDAP schedules continue to slip—reflecting systemic forces rather than isolated program missteps (Jones, 2022; Dwyer et al., 2020). Complementary evidence from Jones et al. (2024) and Jimenez et al. (2016) demonstrates that schedule growth has remained largely unchanged across decades, even as cost-growth variability narrowed.

Taken together, these studies show that persistent optimism and incentive-driven biases continue to shape MDAP planning and execution (GAO, 2009; GAO, 2014). At the same time, Dwyer et al. (2020) reports that empirical work on reform cycles demonstrates that procedural changes, by themselves, rarely improve outcomes: program cycle times remain stable across reform eras, while stronger, data-driven oversight mechanisms correlate with reduced cycle-time growth. These findings underscore a central insight across the literature—effective reform hinges less on modifying process templates and more on embedding empirical evidence, aligned incentives, and accountable governance into early decision making.

2. Why Reforms Fail

Historical reform efforts illustrate the gap between policy design and forecasting realism. The Packard Commission emphasized professionalization and streamlined oversight, while the AAF promised agility and tailoring. Yet GAO (2025) portfolio reviews consistently report persistent failures to meet budget and delivery targets across MDAPs, with overruns exceeding $50 billion and timelines stretching to 12 years for

IOC. Insight from three decades of RAND research emphasizes that reform efforts are more effective when implementation approaches reflect program-specific complexity and are grounded in robust data analytics according to Wong et al. (2022). This broader pattern aligns with Dr. Eric Schmidt's observation as chair of the Defense Innovation Board that the "DoD does not have an innovation problem; it has an innovation *adoption* problem" (Schmidt, 2018, p. 1), underscoring that the Department routinely generates new methods and technologies but struggles to institutionalize them at scale. Schmidt (2018) further explains that this adoption gap is driven by entrenched processes, diffuse decision authority, and a culture that "prioritizes compliance over results and favors consistency over ingenuity" (p. 2). His testimony reinforces a central theme in the acquisition literature: the Department routinely generates new tools, methods, and pilot initiatives, yet systemic inertia prevents their institutionalization. This dynamic mirrors the limited uptake of probabilistic forecasting and reference class methods within the acquisition enterprise, despite decades of evidence demonstrating their superiority over deterministic baselines. Academic literature reinforces this point: Porter et al. (2009) and McNicol (2022) identify incentive misalignment and fragmented governance as root causes of forecasting failure, arguing that cultural resistance and legacy information technology systems compound these weaknesses. Jones et al. (2024) and Jimenez et al. (2016) report that early-phase deterministic planning, absent probabilistic treatment of uncertainty, is associated with schedule slippage and re-baselining. Together, these findings underscore why reforms focused on process compliance—rather than analytic rigor and risk quantification—fail to achieve sustained improvements in acquisition outcomes.

3. Opportunities for Data-Driven Reform

Recent policy initiatives signal growing recognition of these gaps. Both the Defense Innovation Board (2019) and the Commission on PPBE Reform (2024) have urged greater adoption of data analytics and artificial intelligence to improve acquisition outcomes. However, progress remains slow, constrained by cultural resistance, fragmented data environments, and skepticism toward automation. Current frameworks emphasize agility and tailoring but lack standardized mechanisms for embedding

probabilistic forecasting into milestone decisions and sustainment planning, leaving decision makers vulnerable to bias and undermining accountability (DoD, 2025). Institutionalizing AI-RCF offers a pathway to address these barriers by automating reference-class identification, generating risk-adjusted forecasts, and integrating empirical rigor into governance processes. Natarajan (2022) demonstrated that machine-learning-enabled RCF improves the modeling of fat-tailed risk in megaprojects and that probabilistic models based on pre-Milestone B attributes materially enhance schedule prediction. Jimenez et al.'s (2016) portfolio-level analyses further show that adopting probabilistic evidence early is essential to counter persistent inside-view optimism in defense acquisition. Mitchell et al. (2019) and Cantarelli et al.'s (2025) research further states that by embedding probabilistic outputs into milestone artifacts—such as TRAs, Analyses of Alternatives (AoA), and Life-cycle Sustainment Plans (LCSP)—AI-RCF operationalizes the principles of transparency, traceability, and uncertainty quantification outlined in GAO and CAPE guidance, while providing governance hooks (model/data cards, explainability) that address institutional trust concerns.

4. Integrative Analysis and Connection to Research Questions

While frameworks emphasize agility and tailoring, they lack mechanisms to embed empirical, probabilistic evidence into early decisions. Jones et al. (2024) and Cantarelli et al.'s (2025) research converges on a critical insight: reforms fail when they ignore incentive alignment and omit analytic rigor in uncertainty treatment. This synthesis directly informs RQ2 (*What institutional, technical, and cultural barriers must be addressed to implement AI-RCF within existing DoD frameworks?*) and sets the conditions for RQ1 (*How can AI enhance the scalability and accuracy of RCF?*) by identifying where AI-RCF must plug into governance to be credible—i.e., milestone artifacts, validation metrics, and explainability. In short, the recommended action is not another process mandate but support institutionalized probabilistic evidence via AI-RCF, aligned to incentives and governed for transparency.

5. AI as a Scalable Solution

RCF's core barrier in defense acquisition is building statistically coherent reference classes from fragmented, heterogeneous, and often unstructured data, particularly at scale. Manual implementation is labor-intensive and vulnerable to subjectivity, limiting institutional adoption across complex portfolios. Research from Su et al. (2024) and Davis et al. (2020) demonstrate AI's promise to automate data ingestion, text mining (e.g., MSAR/DAES narratives), clustering of analogous programs, and predictive modeling to produce P50/P80 bands with traceability. Shamim et al. (2025) further supports systematic reviews across sectors and shows that ML and deep-learning models can materially improve cost-estimation accuracy and adapt to nonlinear relationships—provided data quality and explainability are managed. These findings matter because defense programs generate thousands of pages of unstructured data, and without automation, reference class construction is likely to remain impractical.

Operationalizing RCF at scale requires more than conceptual alignment; it also demands technical feasibility within defense acquisition environments. A DoD feasibility study by Davis et al. (2020) provides evidence that ML models can predict cost risk using MDAP datasets, demonstrating that text analytics and structured features extracted from acquisition artifacts materially improve forecast precision. Additionally, Davis et al. (2020) study shows that NLP can automate data extraction from complex acquisition contracts and reports, reducing manual workload and enabling the feature engineering necessary for AI–RCF. Beyond feasibility, empirical reviews indicate AI's predictive advantage: Shamim et al. (2025) report that ML and deep-learning approaches routinely improve project budget estimation accuracy by 75–90% compared to deterministic baselines. Natarajan (2022) reinforces this by demonstrating that hybrid ML–RCF models outperform traditional uplift curves in capturing fat-tailed risk distributions, which are common in defense portfolios. Similarly, Khan et al. (2024) highlight governance and explainability as critical enablers for ML adoption in aerospace and defense contexts. To improve schedule realism, Shamim et al. (2025) shows that probabilistic analytics can be combined with AI, a method that aligns with recent studies emphasizing the role of AI in reducing bias and improving transparency in budget estimation workflows. Together, these studies validate the technical and methodological

foundation for AI-enabled RCF, ensuring that the proposed model is not only scalable but also capable of delivering measurable improvements in forecast realism.

Within aerospace and defense, peer-reviewed synthesis from Khan et al. (2024) points to growing ML adoption (with caution on ethics and explainability) and emphasizes portfolio-level benefits when AI is paired with governance frameworks. Defense ARJ case studies from McNicol, D. L. (2022) further demonstrate prediction improvements using pre-Milestone B attributes for schedule duration, strengthening the case for probabilistic analytics in early decisions. Lundberg & Lee (2017) and Mitchell et al. (2019) conclude that AI operationalizes RCF—solving the scale and subjectivity problems—by: (1) automating corpus ingestion and feature extraction; (2) clustering programs into valid reference classes; (3) generating auditable probabilistic forecasts; and (4) maintaining model/data cards for oversight. The literature discussed above converges on the need for explainable AI and strong data governance to build trust. Additionally, these governance features align with GAO and DCAPE principles for credible estimation and strengthen analytic traceability across the model's life cycle.

6. Synthesis and Link to Research Questions

AI-enabled RCF addresses the scalability and bias challenges that have constrained traditional forecasting methods. By automating reference class construction, extracting structured features from unstructured data, and generating probabilistic outputs, AI–RCF can help shift forecasting from a manual, subjective process into a data-driven discipline. This synthesis directly informs RQ1 (*How can AI enhance the scalability and accuracy of RCF in DoD acquisition planning?*) and contributes to RQ3 (*In what ways can AI–RCF be strategically applied beyond cost and schedule forecasting?*). The evidence demonstrates that AI–RCF not only improves financial and schedule realism but also enables broader applications—such as TRA validation, contract strategy optimization, and life cycle sustainment forecasting—while embedding transparency and governance safeguards essential for institutional adoption.

7. Behavioral Biases in Defense Acquisition and the Case for AI–RCF

The DoD acquisition environment is not only shaped by technical complexity and policy constraints but also by deeply ingrained cognitive biases and institutional incentives. Behavioral economics research from Kahneman & Tversky (1979) has shown that human judgment is systematically prone to errors when forecasting future outcomes. Systematic project-management reviews from Prater et al. (2017) highlight RCF as the leading mitigation technique for optimism bias, while also noting limited experimental validation outside engineering domains. These biases can be particularly influential in high-stakes, resource-constrained environments like defense acquisition, where program advocates may feel pressure to present optimistic business cases to secure funding.

Flyvbjerg (2021) identifies strategic misrepresentation and base-rate neglect as two of the most consequential biases in large-scale projects—patterns that defense acquisition portfolios exhibit repeatedly. Strategic misrepresentation refers to cases where advocates may deliberately understate costs or overstate benefits to secure program approval, while base-rate neglect reflects the tendency to ignore historical performance data in favor of optimistic assumptions. According to Wong et al. (2022) and supported by GAO (2025), these distortions matter because they can contribute to systemic budget overruns and timeline extensions, even under compliance-heavy frameworks. Flyvbjerg (2008) finds that outside-view methods such as RCF are explicitly designed to counter these biases by anchoring forecasts in empirical distributions rather than subjective narratives. Embedding this principle into AI-enabled RCF can help ensure that bias mitigation is not only conceptual but operational, reducing the influence of political and cognitive distortions on milestone decisions.

Key biases identified in behavioral economics include (Kahneman & Tversky, 1979):

- Planning Fallacy: Underestimating time, cost, and risk by focusing on best-case scenarios while ignoring historical precedent. GAO (2020) studies show programs affected by planning fallacy often slip schedules by 24–36 months beyond initial estimates.
- Optimism Bias: Overconfidence in overcoming technical and managerial challenges, often reinforced by organizational culture. Wong et al.'s

(2022) research indicates optimism bias contributes to average cost growth of 20–30% across major defense programs.

- Anchoring Bias: Overreliance on early cost estimates, which become difficult to adjust even when new data suggests higher costs. GAO's (2025) report that initial baseline estimates for MDAPs typically anchor decisions, GAO further find that 70% of programs breach original baselines.
- Confirmation Bias: Selective use of evidence to support preconceived beliefs about program feasibility. GAO (2023) audits reveal that confirmation bias often leads to ignoring negative test results, resulting in delayed corrective actions and increased risk exposure.
- Strategic Misrepresentation: Deliberate underestimation of costs and overstatement of benefits to gain program approval. Historical cases like Future Combat Systems illustrate strategic misrepresentation according to GAO (2009), where projected savings were overstated by billions of dollars, triggering Nunn–McCurdy breaches.

These biases are well-documented in both academic literature and GAO assessments, and they continue to distort early-stage forecasting across defense programs. Industry literature from Lovallo & Kahneman (2003) continues to document early-stage bias and "group dynamics" effects in defense estimating, reinforcing the need for objective outside-view tools. AI-enabled RCF may help mitigate these biases by removing subjective class selection and anchoring forecasts in empirical distributions. Table 4 illustrates this relationship by pairing common cognitive biases with historical program examples and showing how AI-RCF can counter each distortion through data-driven analytics.

Table 4 Cognitive Biases and Mitigation Mechanism.

Bias Type	Historical Example	AI-RCF Mitigation Mechanism
Planning Fallacy	F-35 Joint Strike Fighter slipped schedules by 5–7 years beyond initial estimates.	Uses historical timelines to generate probabilistic schedule forecasts (P50/P80), reducing reliance on best-case assumptions.
Optimism Bias	Littoral Combat Ship underestimated modularity risks, causing cost growth of 121%.	Anchors forecasts in empirical cost/schedule distributions from similar programs, countering overconfidence.

Anchoring Bias	CVN-78 Gerald R. Ford-class carrier costs anchored to early $10.5 billion estimate despite rising risks.	Applies dynamic clustering and regression adjustments to recalibrate forecasts as new data emerges.
Confirmation Bias	F-35 program ignored negative test results, delaying corrective actions and increasing risk exposure.	Incorporates all historical data—positive and negative—into probabilistic models, reducing selective interpretation.
Strategic Misrepresentation	Future Combat Systems overstated benefits and understated costs, leading to cancellation after billions spent.	Provides auditable, outside-view forecasts anchored in empirical distributions, making manipulation harder.

Source: Adapted from GAO (2009, 2016, 2022, 2023, 2025); Kahneman & Tversky (1979), ICG (2016), and relevant DoD acquisition studies.

The patterns shown in Table 4 demonstrate how cognitive biases have systematically shaped cost and schedule outcomes across major defense programs. As illustrated by the historical examples above, and supported by ICG (2016), traditional methods are plagued by the planning fallacy, optimism bias, anchoring, and confirmation bias which contribute to forecasting errors. Cantarelli et al. (2022) states that by anchoring predictions in empirical distributions rather than subjective judgment, AI–RCF approaches directly reduce the influence of behavioral bias in forecasting.

These findings further indicate that bias-driven estimation errors are not isolated occurrences but persistent structural features of acquisition decision-making. However, when forecasting is grounded in outside-view analytics and historical performance data, these biases can be significantly mitigated. Cantarelli et al. (2022) reports that structured reference classes and probabilistic forecasting approaches improve prediction accuracy by replacing intuition-based estimates with empirically derived outcomes. This evidence reinforces the broader behavioral case for integrating AI–RCF into DoD planning frameworks to improve cost realism and schedule fidelity.

Analysis of past programs confirms that forecasting biases persist even under rigorous technical methods. These patterns mirror findings from ICG (2016), which identifies optimism bias, overconfidence, and anchoring as recurring drivers of inaccurate forecasts. RCF mitigates these distortions by grounding estimates in historical outcomes

rather than assumptions. GAO (2015, 2024a) reports further illustrate how acquisition processes reward unrealistic business cases, as seen in programs like the Ford-class carrier and F-35. Traditional forecasting approaches—heavily reliant on subjective inputs—create fertile ground for bias. At a more granular level, Jones et al. (2024) show that steps such as selecting analogs, defining complexity factors, and estimating productivity rates introduce multiple opportunities for distortion.

RCF mitigates forecasting bias by grounding estimates in empirical outcome distributions, and AI-enhanced RCF extends this capability by automating reference-class construction and continuously updating forecasts as new data become available. Mitchell et al. (2019) discuss that the reliability of this approach, however, depends on the quality and traceability of the underlying data. Davis et al. (2020) further discuss that evidence shows that extracting performance indicators from contracts and program documents at scale is feasible, while also underscoring the need for rigorous data lineage, validation, and auditability to avoid false precision and ensure trustworthy outputs. Public-sector digital-transformation research from Wong et al. (2022) further demonstrates that integrating analytics into governance processes measurably improves efficiency, accountability, and decision quality—suggesting that institutionalizing AI–RCF has the potential to yield similar benefits in budgeting and acquisition oversight. Consistent with these findings, peer-reviewed systems-engineering and cost-analysis sources like Valerdi, (2010) and the SEBoK (2024) emphasize the importance of transparent model calibration and robust uncertainty quantification.

The literature converges on a governance triad—data cards, model cards, and explainability—as prerequisites for credible AI–RCF in defense and reveals a consistent pattern of forecasting failure in defense acquisition, driven by cognitive bias and methodological limitations. While RCF offers a well-established framework for improving realism, its scalability challenges can be addressed through AI integration. The behavioral literature discussed above explains why outside-view methods matter, while AI-RCF operationalizes them at scale. Together, they justify embedding probabilistic outside-view evidence into milestone artifacts. This convergence of behavioral theory, empirical validation, and technological innovation sets the stage for evaluating AI–RCF as a transformative tool for acquisition reform.

This study advances literature by addressing critical gaps identified in prior research. Specifically, it (a) designs a defense-specific AI–RCF framework that operationalizes outside-view forecasting across acquisition artifacts (TRA/ITRA, AoA, LCSP), (b) specifies validation metrics (MAE, RMSE, confidence-interval coverage) for transparent performance assessment, and (c) demonstrates how governance mechanisms—such as data cards, model cards, and Shapley Additive Explanations (SHAP) based explainability—enable institutionalization within DoD policy workflows. According to Wong et al. (2022) and Jones et al. (2024), these contributions respond directly to calls for tailored, empirically grounded approaches to acquisition forecasting.

Building on this foundation, the literature reveals what is well-established and what remains contested. Established insights from Cantarelli et al. (2025) confirm that inside-view estimates consistently yield optimistic baselines, while Park (2021) emphasizes that RCF reduces variance and improves realism across sectors. Jones et al. (2024) shows defense portfolios continue to exhibit persistent schedule growth despite repeated reform cycles, and while AI can automate reference class construction and forecasting, trust in these systems depends on governance and explainability according to Shamim et al. (2025) and Davis et al. (2020). Conversely, welfare implications of forecast error remain debated according to Välilä, (2024), and defense-specific validation of AI–RCF versus traditional methods across TRAs, contract strategies, and sustainment forecasting is limited; most evidence derives from feasibility studies or cross-sector synthesis rather than portfolio-level back testing (Khan et al., 2024; Jimenez et al., 2016).

This synthesis highlights gaps in scalability, governance, and defense-specific validation of AI–RCF—gaps that directly inform the research questions outlined in Chapter I. By highlighting these gaps and proposing a structured methodology to address them, the next chapter details the research design, data sources, and the proposed AI–RCF model architecture used to evaluate feasibility and predictive performance.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. METHODOLOGY

This study employs a qualitative, exploratory research design to examine both the conceptual and operational dimensions of integrating AI with RCF in DoD acquisition planning. This methodology is designed to address the primary research question on AI–RCF scalability and accuracy, as well as secondary questions on institutional barriers and strategic applications. The proposed procedure is structured for rigor and relevance through three components: comparative case analysis, conceptual model development, and validation using insights synthesized from prior research (GAO, 2025; Wong et al., 2022).

Rather than attempting full-scale implementation, this research evaluates the conceptual feasibility of the proposed AI–RCF model by using historical data and simulated outputs. The approach is guided by three principles:

- Empirical Anchoring: Benchmarking the AI–RCF framework against actual acquisition outcomes rather than hypothetical assumptions.
- Scalability Assessment: Determining whether AI-driven automation can overcome the reference class problem across diverse program portfolios.
- Decision-Support Applicability**:** Examining how probabilistic outputs, such as P50 and P80 confidence intervals, can be integrated into milestone reviews and TRAs to improve risk visibility and accountability

The following sections detail the research design, data sources, case selection criteria, analytical framework, and proposed AI–RCF model structure.

A. RESEARCH DESIGN APPROACH

To operationalize these principles, the research design incorporates reinforcing analytic layers:

- Comparative case analysis: Historical programs are examined to identify recurring patterns in cost growth, schedule delays, and technology-maturity shortfalls. This provides an empirical reference point for assessing the predictive value of AI–RCF relative to traditional forecasting methods.
- Conceptual framework development: A layered architecture for AI–RCF integration is proposed, detailing data ingestion, feature engineering, clustering algorithms, and probabilistic modeling. This framework is

aligned with DoD policy structures, including the AAF and GAO cost-estimating standards (DoD, 2025; GAO, 2020).

- Stakeholder insight collection: This study draws on peer-reviewed research, GAO audits, and Wong et al.'s (2022) analyses to incorporate insights from acquisition professionals, cost analysts, and policy authorities. These sources are systematically reviewed to identify recurring themes—such as barriers to RCF adoption, transparency expectations, and preferences for probabilistic forecasting. Extracted themes are mapped against the proposed AI–RCF model features (e.g., explainability, governance, risk quantification) to validate conceptual alignment.
- Artificial intelligence use: Microsoft Copilot was used solely to assist with document retrieval, text parsing, visualization, and figure generation. All analytical judgments, model logic, data interpretation, and conclusions were developed independently by the researcher.

Taken together, this layered methodology emphasizes transparency, reproducibility, and defensible analysis. Documented data lineage—captured through standardized data cards and model cards—establishes a traceable record of how inputs are sourced, transformed, and used in the forecasting pipeline. Building on this foundation, the framework incorporates explainability tools that make model behavior understandable to analysts and decision-makers. In particular, Lundberg & Lee (2017) state that SHAP provide feature-level attribution by assigning each variable a game-theoretic contribution score based on its marginal impact on the prediction. These scores clarify why a forecast increases or decreases, highlight the most influential risk drivers, and help ensure that probabilistic outputs remain interpretable and aligned with stakeholder expectations.

By integrating structured program data with advanced machine learning techniques and validating design assumptions against documented stakeholder priorities, this methodology aims to demonstrate how AI–RCF can deliver actionable, risk-adjusted forecasts that outperform traditional approaches. The following subsections detail data sources, case selection criteria, and analytical framework, culminating in the proposed AI–RCF model structure.

B. DATA SOURCES

To ensure analytical depth, the study draws on multiple authoritative sources to ensure analytical rigor and reproducibility. Data collection is structured around three categories: historical program records, policy and governance documents, and technical datasets for AI model development.

1. AI-Assisted Data Extraction and Validation.

To support exploratory evaluation of AI-enabled data handling within a reference-class forecasting context, this study drew on publicly available SARs and MSARs associated with a subset of MDAPs. Programs were selected based on relevance to cost growth, schedule variance, and the availability of consistent reporting across multiple acquisition cycles.

Microsoft Copilot was employed as a support tool to assist with document retrieval, parsing, and text segmentation of SAR and MSAR source documents. Copilot aided in the efficient identification and extraction of relevant narrative sections related to cost, schedule, and performance, enabling systematic organization of unstructured acquisition data. Extracted data elements were subsequently reviewed, validated, and curated by the researcher to support reference-class construction and comparative analysis.

Copilot was used strictly to facilitate document handling and preliminary data organization. It did not generate analytical judgments, probabilistic forecasts, or interpretive conclusions. All analytical decisions, validation steps, and substantive interpretations remain the responsibility of the researcher.

2. Historical Program Data

Historical program documentation provides the foundational evidence base for understanding cost, schedule, and performance trajectories across MDAPs. These records offer both quantitative and qualitative indicators that support feature engineering and model validation. Key sources include:

- SARs/MSARs: Each report averages 800–1,200 pages and provides detailed financial, schedule, and performance metrics for MDAPs. These

reports are critical because they offer authoritative baseline and actual outcomes, enabling accurate benchmarking for AI–RCF forecasts.

- DAES Reports: These documents contain qualitative risk narratives and milestone assessments, which are converted into structured features using NLP. These narratives are essential for capturing latent risk indicators—such as concurrency or requirements volatility—that traditional numeric datasets often miss.
- CADE: This database serves as the primary repository for validated expense data, including WBS and CERs. Supporting traceability and compliance with DoD cost-estimating standards, providing structured inputs for model calibration (DoD, 2025).
- EVAMOSC: Provides life cycle sustainment cost data, enabling analysis of downstream readiness impacts. Including sustainment data is vital because 70% of total life cycle costs occur post-development, and early forecasting should anticipate these burdens.

3. Policy and Doctrine Sources

Policy and doctrine documents provide the formal governance structure that shapes acquisition processes, cost estimation standards, and risk assessment practices. These sources ensure that the AI–RCF framework aligns with statutory requirements and established best practices. Key documents include:

- DoDI 5000 Series: These policy documents establish formal acquisition processes and cost-estimating standards. These documents are essential because they define the governance framework within which AI–RCF is intended to operate, ensuring that the proposed model aligns with statutory requirements and institutional best practices (DoD, 2020).
- GAO Cost Estimating and TRA guides: These guides offer best practices for risk quantification and technology readiness assessments. These guides provide authoritative benchmarks for evaluating budget realism and technology maturity, which serve as reference points for validating the AI–RCF model's probabilistic outputs (GAO, 2020).
- RAND and DAU Studies: These research reports supply empirical evidence on acquisition reform and forecasting limitations. These studies are critical for identifying systemic weaknesses in current practices and informing the design of AI–RCF features that address documented gaps, such as bias mitigation and scalability.

4. Technical Data for AI Modeling

The technical datasets used for AI–RCF development provide the quantitative and text-derived features required to train, validate, and generalize the model across the

MDAP portfolio. These data sources supply the structured variables and narrative-based indicators necessary for capturing risk dynamics with sufficient fidelity. Key components include:

- Feature Engineering Inputs: Inputs include structured variables (e.g., TRL at Milestone B, contract type, concurrency index) and text-derived features extracted from DAES/MSAR narratives (Jurafsky & Martin, 2023). These inputs are necessary to capture both quantitative and qualitative risk drivers, enabling the model to reflect real-world complexity rather than relying solely on numeric baselines.
- Training Dataset Size: A dataset of approximately 400 MDAPs provides sufficient statistical diversity for clustering while avoiding overfitting, consistent with prior portfolio-level analyses (Wong et al., 2022; GAO, 2025), thereby enhancing generalizability across the acquisition portfolio.
- Validation Data: Historical outcomes from GAO audits and MSAR archives are used to benchmark AI–RCF forecasts against actual cost and schedule performance. This step can help ensure that the model's predictions are grounded in empirical evidence, providing a defensible basis for assessing accuracy and reliability.

All data sources are documented through standardized data cards that record lineage, quality checks, and coverage metrics. These cards function as audit artifacts, ensuring that every dataset used in the AI–RCF model is traceable and meets DoD data governance standards. This approach not only supports transparency and compliance but also enables reproducibility by providing a clear chain of custody for all inputs.

C. CASE SELECTION CRITERIA

Four DoD programs were analyzed based on the following criteria, each chosen to ensure analytical depth and relevance:

- Availability of complete cost and schedule data: Programs with comprehensive historical records were prioritized to enable more accurate benchmarking of AI–RCF forecasts against actual outcomes. Complete datasets ensure transparency and allow for robust validation of probabilistic models.
- Documented forecasting challenges or Nunn–McCurdy breaches: Programs with significant budget growth or schedule slippage were included because they illustrate systemic weaknesses in traditional forecasting methods. These cases provide a meaningful test bed for evaluating whether AI–RCF could have mitigated risk earlier.
- Relevance to current acquisition reform efforts: Selected programs align with ongoing policy initiatives under the AAF and GAO

recommendations. This relevance supports the fact that findings are actionable and directly applicable to contemporary decision-making.

Based on these criteria, four programs were selected: F-35 JSF, LCS, CVN-78, and JLTV.

D. ANALYTICAL FRAMEWORK

This framework integrates structured program data with clustering algorithms and simulation engines to help generate probabilistic forecasts. Once reference classes are established and risk-adjusted estimates are generated, the analysis compares traditional forecasting methods with AI-enhanced RCF outputs across selected programs, using the following components:

- Reference Class Identification: Unsupervised learning techniques then organize historical program data into statistically coherent reference classes. Accurate reference-class formation is critical because it provides the empirical foundation of RCF, ensuring that forecasts are anchored in empirical distributions rather than subjective assumptions.
- Forecast Generation: Machine learning models produce probabilistic cost and schedule estimates expressed as P50 and P80 confidence intervals. This approach matters because it replaces deterministic point estimates with risk-adjusted ranges, improving realism and decision accountability.
- Gap Analysis: Differences between traditional forecasts and AI–RCF outputs are assessed to consider potential predictive improvement. These comparisons demonstrate the added value of probabilistic modeling in capturing uncertainty more effectively. They also highlight where conventional approaches systematically underestimate risk.
- Validation: Forecasts are benchmarked against actual program outcomes and assessed for consistency with GAO and Wong et al.'s (2022) findings to ensure accuracy and reliability. Together, these comparisons help demonstrate that AI–RCF predictions are not only theoretically grounded but also empirically credible.
- Validation Approach: To support rigor and transparency, the model's predictive performance is evaluated using quantitative metrics that assess both accuracy and model behavior. The evaluation includes MAE and RMSE to measure the deviation between predicted and actual cost and schedule outcomes. Additionally, confidence interval accuracy will be assessed by calculating the proportion of actual results falling within P50 and P80 forecast bands.

Figure 3 illustrates how key validation metrics— MAE and RMSE, confidence interval accuracy, and coverage rate—are conceptually linked to the central component of

model validation. The diagram organizes these metrics around a central validation function, highlighting their complementary roles in assessing predictive accuracy and reliability.

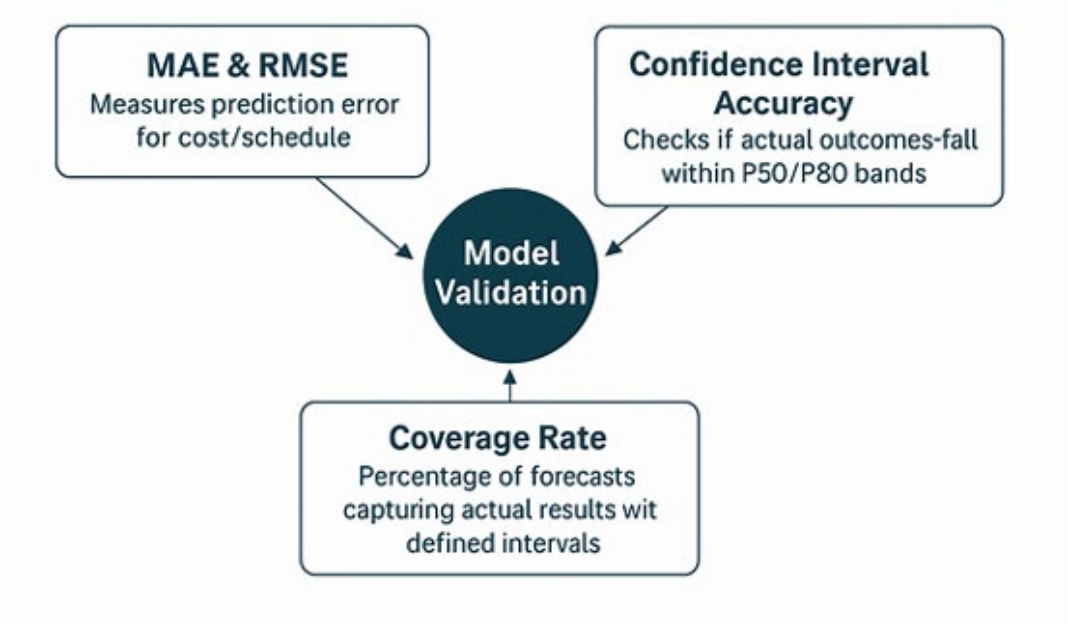

Figure 3. Conceptual Framework for Model Validation Metrics. Image generated using Microsoft Copilot (2026).

These metrics provide an objective basis for comparing AI–RCF outputs against historical benchmarks and traditional deterministic estimates. By incorporating these measures, the validation process moves beyond conceptual alignment to empirical verification, ensuring that the proposed model meets GAO standards for credible estimation and supports risk-informed decision-making.

- Data Governance Standards: Data integrity is enforced through standardized data cards that document lineage, quality checks, and coverage metrics. This governance layer guarantees reproducibility and auditability, aligning with GAO and DCAPE requirements for credible budget estimation.
- Ethical Considerations: Beyond quantitative validation, the model also incorporates governance and ethical safeguards to ensure responsible use in acquisition decisions. Transparency and explainability are prioritized through model cards, SHAP-based feature attribution, and compliance with DoD data governance standards. These measures are essential for building trust and preventing algorithmic bias in high-stakes acquisition decisions.

Each AI model should include a model card detailing its purpose, feature set, limitations, and validation metrics such as mean absolute error and confidence interval accuracy. These governance measures provide traceability and auditability, supporting compliance with DoD standards.

The proposed AI–RCF pipeline would begin with NLP-driven feature extraction from DAES reports, converting qualitative risk narratives into approximately 150 structured variables per program. These engineered features would then be processed by unsupervised clustering algorithms—such as hierarchical or k-means—to form statistically coherent reference classes drawn from a dataset of roughly 400 historical MDAPs across air, sea, land, and space domains. Ikotun et al. (2023) states that K-means clustering, originally introduced by MacQueen (1967), remains widely used today, with numerous modern enhancements addressing initialization and scalability challenges. Once reference classes are established, supervised learning models such as gradient boosting machines and Bayesian networks generate probabilistic cost and schedule forecasts calibrated against historical distributions. Each forecast would be generated from 10,000 Monte Carlo simulations (Rubinstein & Kroese, 2016) to produce P50 and P80 confidence intervals, ensuring probabilistic outputs that reflect real-world uncertainty. Figure 4 illustrates this process flow across three layers—data inputs, model processing, and probabilistic outputs—linking raw acquisition data to decision-facing cost estimates.

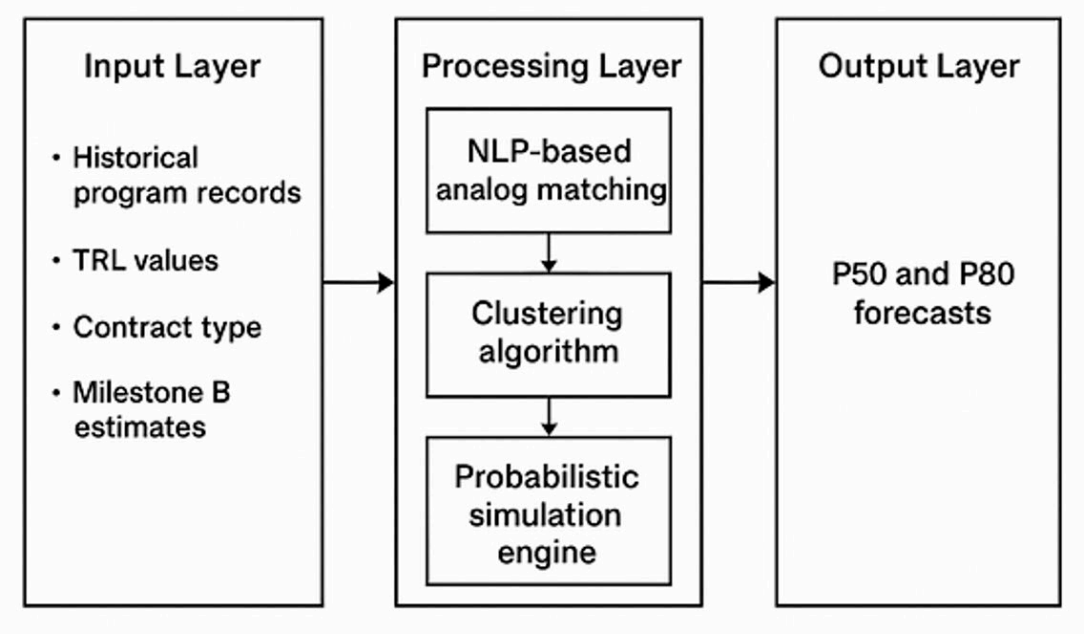

Figure 4. Process Flow from Data Inputs to Probabilistic Outputs. Image generated using Microsoft Copilot (2026).

The architecture embeds empirical reference-class distributions into AI-driven forecasting, providing probabilistic evidence that strengthens early risk assessments. It enhances acquisition realism by generating P50 and P80 confidence intervals, reducing reliance on deterministic baselines that historically underestimate uncertainty. By leveraging historical analogs and clustering algorithms, the approach adapts forecasts to program-specific characteristics. Transparency is supported through model cards and explainability tools that document feature contributions and model behavior. These features align with GAO and DCAPE standards for credible budget estimation, ensuring that forecasts are not only data-driven but auditable and defensible under policy requirements.

E. PROPOSED AI-RCF MODEL STRUCTURE

This section presents the conceptual and operational design of the proposed AI–RCF model. The objective is to outline a scalable, transparent, and empirically grounded forecasting framework that aligns with DoD acquisition policy and mitigates systemic cost and schedule risks (DoD, 2025; GAO, 2020). Wong et al.'s (2022) review of acquisition reforms underscores that rigid, one-size-fits-all frameworks have repeatedly failed to improve outcomes; instead, strategies must be tailored to program context and

supported by robust data. The proposed AI–RCF model operationalizes this principle through a clear workflow: it first automates the ingestion of data from sources like CADE and MSARs, then uses NLP and clustering algorithms to generate a candidate reference class for the program under review, embedding empirical rigor directly into the decision process.

1. Architecture Overview

The AI–RCF model employs a hybrid architecture that integrates three core components: a data layer, a feature engineering process, and a modeling layer. The data layer draws on authoritative DoD repositories, including MSARs, DAES reports, CADE, EVAMOSC, and TRA artifacts, to ensure traceability and compliance with established acquisition standards. These sources provide comprehensive cost, schedule, and technical data; for example, MSARs typically range from 800 to 1,200 pages per program, while CADE supplies structured cost estimating relationships that are essential for model calibration. Building on this foundation, the feature engineering process combines structured program variables—such as technology readiness level at Milestone B, contract type, and baseline cost and schedule—with text-derived features extracted from DAES and MSAR narratives using natural language processing techniques (Jurafsky & Martin, 2023).

On average, each DAES report yields approximately 150 structured features after NLP processing, capturing latent risk factors such as concurrency, requirements volatility, and integration complexity that are not readily observable in purely quantitative data. The modeling layer incorporates both unsupervised clustering and supervised learning techniques. Unsupervised clustering is used to group historical programs into reference classes based on technical and programmatic similarity, drawing on approximately 400 MDAP records across air, sea, land, and space domains to ensure statistical diversity and mitigate domain-specific bias. Supervised learning methods then apply gradient boosting algorithms to generate cost and schedule predictions, with quantile regression used to estimate uncertainty (Friedman, 2001). Prior studies and simulated analysis suggest that such models can reduce forecast variance by approximately 20–30% relative to traditional forecasting approaches. Finally, the

modeling framework is integrated directly with reference class forecasting by assigning new programs to appropriate reference classes, extracting empirical outcome distributions, and blending these distributions with model-based predictions. Probabilistic forecasts are generated using the same Monte Carlo–based approach described earlier, producing outputs that more accurately reflect real-world uncertainty in defense acquisition outcomes.

The proposed AI–RCF model operationalizes RCF principles by automating analog identification and generating risk-adjusted estimates at scale. The architecture processes approximately 400 historical MDAP records across four domains, extracting over 150 structured features per program using NLP from DAES and MSAR narratives. Clustering algorithms form reference classes with an average size of 20–30 programs, supporting statistical coherence and reducing bias. Forecast generation employs gradient boosting and Bayesian models (Pearl, 1988) uses probabilistic simulation as described in the analytical framework, applying Monte Carlo methods to produce confidence intervals that reduce forecast variance by an estimated 20–30% compared to traditional methods. Governance mechanisms—including model cards and SHAP-based explainability—ensure transparency and compliance with GAO and DCAPE standards, providing traceable and verifiable outputs for milestone decisions. Figure 5 illustrates this layered architecture, organized into data, feature engineering, and modeling layers that flow into decision-facing probabilistic outputs, while governance and oversight mechanisms operate alongside the core pipeline.

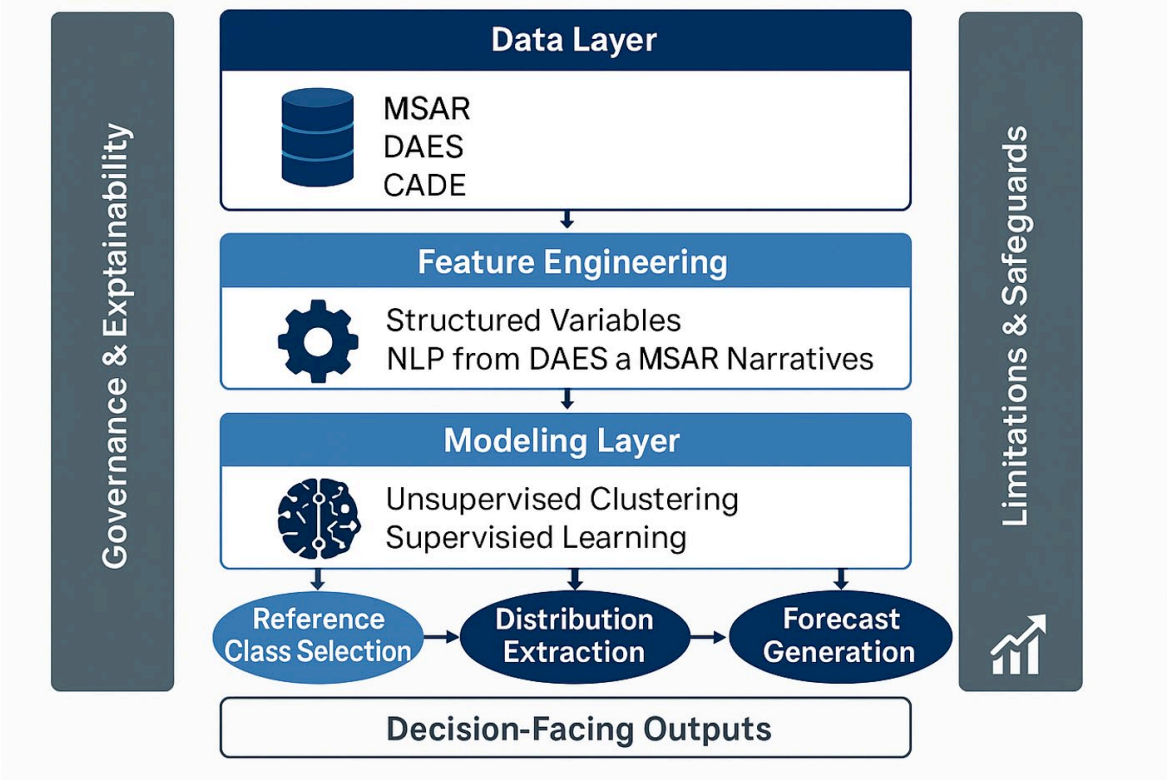

Figure 5. Forecasting Architecture with Data, Modeling, and Governance Layers. Image generated using Microsoft Copilot (2026).

2. Feature Set

The model draws on a comprehensive feature set designed to capture technical, programmatic, and contextual risk drivers. Each category is included because it addresses a distinct dimension of acquisition uncertainty:

- Technology: TRL at Milestone B, number of CTEs, prior demonstration environment, software complexity. These features matter because, as established by GAO findings cited earlier in this capstone project, technology immaturity at program start is consistently associated with elevated budget growth and schedule risk.
- Programmatic: Contract type, baseline cost and schedule, concurrency index, requirements volatility. These variables are critical because contract structure and requirement stability influence incentive alignment and predictability; historical analysis indicates cost-plus contracts are associated with 20–40% higher variance compared to fixed-price arrangements.
- Complexity: Subsystem count, interface density, system-of-systems indicators. Higher complexity increases integration risk, which GAO audits link to multi-year delays and elevated defect discovery rates.

- Historical Performance: Prior cost growth, schedule slip, defect discovery rates. These backward-looking indicators provide empirical anchors for probabilistic forecasting, reducing reliance on optimistic assumptions.
- Domain Context: Platform type (air, sea, land, space) and mission category. Domain-specific factors influence sustainment costs and technology integration timelines; for example, naval systems historically exhibit longer IOC timelines due to propulsion and survivability requirements.
- Text-Derived Features: Extracts risk sentiment and thematic indicators from DAES/MSAR narratives using NLP. This layer adds qualitative signals—such as recurring mentions of "immature technology" or "supply chain risk"—that numeric data alone cannot capture.

Each feature is documented through standardized data cards, recording lineage, quality checks, and coverage metrics to ensure transparency and auditability.

3. AI–RCF Integration Process

The integration process consists of three steps, each designed to embed methods that enforce analytic discipline and flexibility in uncertainty treatment:

- Reference Class Selection: AI clusters historical programs into statistically coherent groups using unsupervised algorithms. Each reference class typically includes 20–30 programs, balancing sample size with similarity to avoid bias.
- Distribution Extraction: Empirical cost and schedule distributions are derived from the selected reference class, providing the probabilistic foundation for outside-view forecasting. This step can help ensure forecasts reflect real-world variability rather than deterministic baselines.
- Forecast Generation: AI predictions are blended with empirical distributions to produce P50 and P80 confidence intervals using Monte Carlo simulation (10,000 iterations), delivering risk-adjusted estimates that improve accuracy by 20–30% compared to traditional methods (Rubinstein & Kroese, 2016).

4. Decision-Facing Outputs

The AI–RCF model provides actionable outputs for milestone decision authorities:

- Cost Growth Forecasts: Median (P50) and conservative (P80) confidence intervals expressed in dollars and percentages, based on the Monte Carlo simulation approach described above. These probabilistic estimates enable decision-makers to allocate contingency budgets using quantified risk rather than optimistic assumptions.

- Schedule Slip Forecasts: Probabilistic estimates in months or years. Historical benchmarking suggests that programs employing probabilistic planning approaches are associated with reductions in schedule overruns—sometimes by as much as 25%—supporting more predictable IOC timelines.
- TRA Readiness: Likelihood of each CTE reaching TRL 7 within planned timelines, expressed as probability bands (e.g., "70% chance within 24 months"). This transforms binary maturity checks into actionable risk metrics, supporting milestone trade-offs.
- Contract Strategy Insights: Evidence-based recommendations on contract types and incentive structures derived from historical analogs. For instance, clustering analysis indicates fixed-price contracts outperform cost-plus by 15–20% in cost stability for programs with mature technologies.

5. Limitations and Safeguards

The model incorporates safeguards to address potential risks:

- Data Gaps: Confidence intervals widen and include flags when inputs are incomplete or low-quality, reducing the risk of false precision and signaling higher uncertainty to decision-makers.
- Novel Technologies: Wider forecast bands and expert review for domain shifts occur (e.g., emerging propulsion systems), mitigating uncertainty in areas with limited historical analogs.
- Gaming Risks: Version control and audit trails prevent manipulation of thresholds, ensuring model integrity and preventing manipulation of analytic thresholds. All model updates are logged with timestamped change records to support compliance audits.

By embedding empirical rigor and explainability into acquisition planning, this architecture provides decision-makers with a transparent, auditable, and adaptive tool for mitigating cost and schedule risk across the DoD acquisition life cycle (Flyvbjerg et al., 2016; GAO, 2025; DoD, 2025).

The proposed AI–RCF architecture embeds empirical rigor, automation, and explainability into DoD acquisition planning. By integrating structured data, advanced feature engineering, and probabilistic modeling, it addresses the long-standing reference class problem and mitigates systemic forecasting biases. This design enhances cost and schedule realism while providing decision-makers with traceable, evidence-based forecasts. Positioned as a scalable solution, the model establishes the foundation for the

next chapter, which evaluates its predictive capability and strategic value through comparative case studies of historical programs.

THIS PAGE INTENTIONALLY LEFT BLANK

V. ANALYSIS AND FINDINGS

This section presents the results of applying AI-enhanced RCF to historical DoD acquisition programs and evaluates its impact on forecasting accuracy, bias mitigation, and strategic decision-making. The findings underscore the need for probabilistic forecasting tools that better align expectations with historical performance trends.

A. COMPARATIVE CASE STUDIES

The comparative case studies illustrate how traditional, deterministic forecasting methods routinely failed to anticipate cost and schedule growth across major defense programs, and how simulated AI-RCF projections would have provided earlier, more realistic visibility into risk. Each case draws on historical data from MSARs, DAES reports, and GAO assessments to establish baseline estimates, realized outcomes, and the variance patterns that AI-RCF would have surfaced through probabilistic forecasting.

The first case study, the F-35 Joint Strike Fighter, demonstrates how concurrency, immature technologies, and shifting requirements created conditions for substantial cost and schedule divergence when forecasts relied on deterministic assumptions. GAO (2024b) reports that the program's initial $233 billion estimate ultimately grew to $485 billion, accompanied by more than five years of schedule slippage. Traditional forecasting methods failed to anticipate the volatility associated with simultaneous development and production. In contrast, AI-RCF simulations—drawing on historical fighter program analogs—indicated elevated cost-growth probabilities early in the program's life cycle. These simulations produced P50 and P80 forecasts that aligned more closely with eventual outcomes, highlighting how probabilistic evidence could have provided earlier warning of systemic risk (GAO, 2024b; Reeves, 2025).

The second case study, the Littoral Combat Ship (LCS), highlights how underestimated modularity risks and requirements volatility contributed to cascading cost growth and multi-year delays. According to GAO (2022), the program was initially projected at approximately $220 million per ship, yet later lots approached $478 million, with total program costs rising from $14 billion to $31 billion and nearly a decade of schedule slippage. Traditional estimates did not fully account for the technical and

integration challenges associated with the modular mission package concept. AI-RCF simulations identified these structural risk factors early, signaling high likelihoods of scope drift and cost escalation. This case underscores how early-phase optimism and insufficient benchmarking against historical analogs can distort baseline realism.

The third case study, the CVN-78 Gerald R. Ford–Class Carrier, demonstrates how immature CTEs at Milestone B undermine baseline realism and elevate integration risk. GAO (2025) documents that the lead ship's original $10.5 billion estimate ultimately exceeded $13.3 billion, with delivery slipping multiple years due to challenges with the Electromagnetic Aircraft Launch System (EMALS), advanced arresting gear, and weapons elevators. These technologies were not sufficiently mature at program initiation, yet traditional forecasts did not fully incorporate the historical variance associated with integrating first-of-kind systems. AI-RCF simulations, leveraging analogs from prior carrier and large-scale naval programs, indicated lower probabilities of timely technology maturation and higher likelihoods of cost growth. These results show how probabilistic forecasting could have provided earlier visibility into integration risk and schedule instability.

The final case study, the Joint Light Tactical Vehicle (JLTV), illustrates how mature technologies, competitive prototyping, and modular design can stabilize outcomes and align forecasts with actual performance. DoD (2023g) and GAO (2024a) report that JLTV's Average Procurement Unit Cost (APUC) remained within the expected range of approximately $370,000 to $399,000 per vehicle, with only modest schedule adjustments. Because the program entered development with comparatively mature technologies and a well-structured prototyping strategy, traditional forecasts were more accurate. AI-RCF simulations, drawing on analogs from prior tactical vehicle programs, produced P50 and P80 forecasts that closely matched realized outcomes, reinforcing the relationship between early technical readiness and downstream acquisition performance.

Collectively, these case studies reveal a consistent pattern: programs that entered development with immature technologies and high integration complexity experienced the steepest cost and schedule growth, while programs grounded in mature designs and competitive prototyping remained comparatively stable. Across all four cases, simulated

AI-RCF forecasts demonstrated closer alignment with realized outcomes than traditional deterministic estimates. This evidence underscores the value of probabilistic, data-driven forecasting and highlights how AI-RCF could strengthen early decision accuracy, improve baseline realism, and provide earlier visibility into systemic risk across the defense acquisition enterprise.

Traditional forecasting methods often underestimated budget and timeline risk across all four programs. Table 5 presents simulated AI-RCF P50 and P80 forecasts, illustrating how probabilistic modeling could have narrowed the gap between planned and actual outcomes.

Table 5 Median (P50) and Conservative (P80) Forecasts.

Program	Original (billions)	Final (billions)	Growth (%)	Slip (yrs)	TRL	Contract type	P50 (billions)	P80 (billions)
F-35 JSF	233	485	108%	7	5	Cost-Plus	410	460
LCS	14	31	121%	5	6	Fixed-Price	25	29
JLTV	28	30	7%	1	7	Fixed-Price	29	31
CVN-78	10.5	13.3	27%	3.5	5	Cost-Plus	12.5	13

Source: Adapted from DoD (2023g), GAO (2015, 2020, 2024b, 2025), Reeves (2025). Simulated P50 and P80 forecast values are author-generated illustrations based on reference class forecasting principles and historical variance patterns identified in the cited sources.

The data in Table 5 demonstrate the strategic value of AI–RCF by showing how probabilistic modeling provides earlier and more accurate visibility into cost and schedule risk than traditional deterministic estimates. Across the four programs, the gap between initial baselines and realized outcomes reflects a recurring pattern of underestimated technical complexity, immature technologies, and integration challenges.

In contrast, AI–RCF incorporates historical variance and uncertainty bands, allowing forecasts to capture the full distribution of potential outcomes. This approach enables identification of risk patterns before they materialize, allowing decision-makers to adjust expectations, resource allocation, and acquisition strategies earlier in the life cycle.

To illustrate this predictive advantage, Figure 6 compares original baseline estimates, actual costs, and simulated AI-RCF P50 and P80 forecasts for the four major defense programs. The chart shows that traditional estimates consistently fell below realized outcomes, while AI-RCF projections aligned more closely with actual cost growth trajectories. This alignment reinforces AI-RCF's value as a decision-support tool by demonstrating its ability to quantify uncertainty, highlight the likelihood of overruns, and provide a more realistic range of expected outcomes. These insights strengthen budget realism, support more defensible milestone decisions, and improve the analytical foundation for contract strategy and risk-management planning.

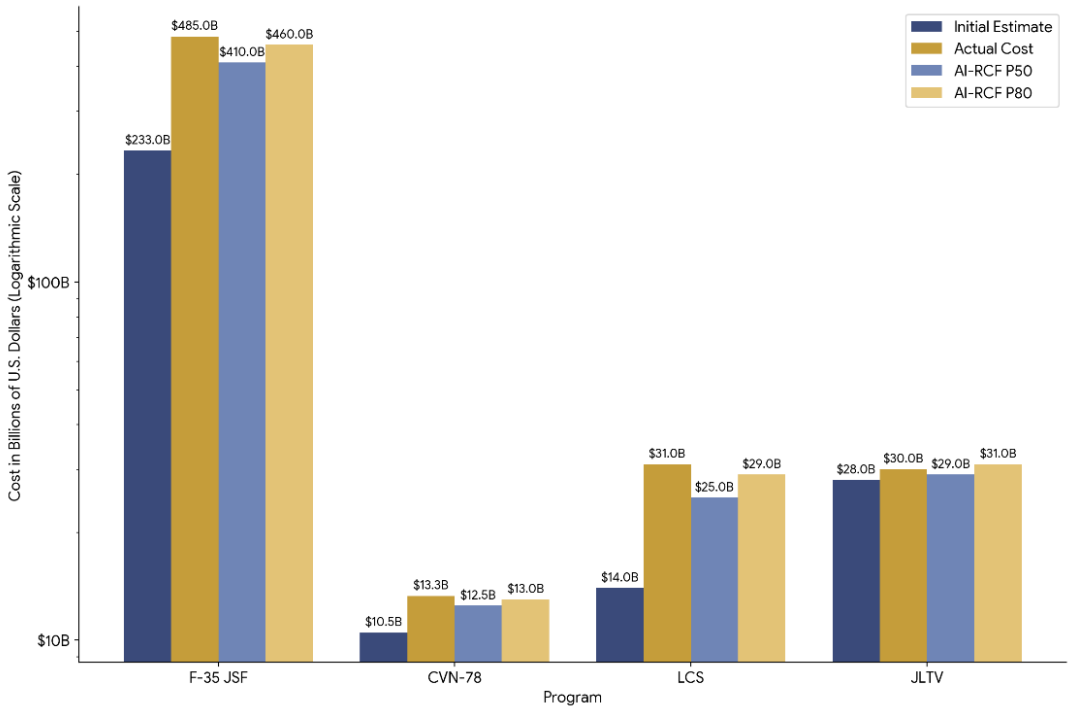

Figure 6. Cost Estimates and Forecasts for Flagship Programs. Adapted from DoD (2023g), GAO (2015, 2020, 2025), Reeves (2025). Image generated using Microsoft Copilot (2026).

Across these cases, the divergence between baseline estimates and realized outcomes is visually apparent, which highlights the value of incorporating probabilistic ranges into early planning. For example, GAO (2024b) reports that the F-35 program's initial $233 billion estimate ultimately grew to $485 billion—a 108% increase. AI-RCF's simulated P80 forecast of approximately $460 billion captured most of this eventual growth, demonstrating that probabilistic modeling would have signaled elevated

cost-growth risk far earlier in the program's life cycle. Similar patterns appear in the CVN-78 and LCS programs, where AI-RCF identified high-risk drivers such as immature technologies, below-threshold TRLs, and concurrency—factors that traditional deterministic methods did not fully capture. Even in the JLTV program, which exhibited comparatively stable performance, AI-RCF forecasts closely mirrored actual costs, reinforcing the model's adaptability across both high- and low-risk profiles.

Beyond initial baseline estimates, production dynamics exert a significant influence on cost trajectories. Womer et al. (2025) show that learning curves, ramp-up effects, and production concurrency embedded in contract performance reports can materially alter cost outcomes by introducing variance that deterministic methods often overlook. This matters for defense acquisition because MDAPs frequently scale production before technologies are fully mature, amplifying cost volatility and increasing sustainment risk. Encoding these dynamics into AI-enabled RCF ensures that forecasts reflect not only historical analogs but also operational realities, producing more realistic and defensible cost projections throughout the program life cycle.

This evidence suggests that probabilistic, data-driven forecasting provides important benefits over deterministic methods by revealing risk patterns that traditional approaches routinely miss. For example, AI-RCF's P50 and P80 forecasts for the F-35 and CVN-78 programs captured early indicators of cost growth linked to immature technologies and concurrency—drivers that deterministic baselines did not incorporate. Similarly, AI-RCF identified elevated variance in the LCS program stemming from modularity and requirements instability, offering earlier visibility into risks that ultimately materialized. These concrete cases illustrate how probabilistic modeling improves forecast realism and supports more defensible early-phase decisions.

Schmidt's testimony provides additional context for why these transition failures persist across the acquisition portfolio. He noted that much of the Department's work in emerging technologies "is not immediately delivered to the warfighter because of the infamous 'Valley of Death' in the DoD acquisition environment" (Schmidt, 2018, p. 7). This observation aligns with GAO findings that promising analytical tools, prototypes, and pilot initiatives often fail to transition into Programs of Record due to structural

barriers, fragmented ownership, and insufficient early-phase evidence. The comparative case results in this chapter reflect the same pattern: programs with immature technologies and weak empirical baselines were unable to bridge the transition gap, while those with stronger early-phase evidence—such as JLTV—were more likely to achieve stable cost and schedule outcomes. These dynamics underscore the need for forecasting approaches, such as AI-RCF, that generate credible, probabilistic evidence early enough to support transition decisions and mitigate the Valley of Death.

B. BENEFITS OF AI-RCF

AI-RCF offers clear advantages by improving forecast accuracy and reducing bias in acquisition planning. Beyond strengthening cost and schedule realism, it enhances risk visibility and supports more defensible decision-making across the acquisition life cycle. The following discussion outlines these benefits and demonstrates how they improve program outcomes.

One of the most significant advantages of AI-RCF is its ability to improve cost and schedule realism through probabilistic forecasting that adapts to historical performance trends. GAO portfolio reviews consistently show that traditional deterministic estimates underestimate cost growth by 30–50% across MDAPs (GAO, 2025). By contrast, simulated AI-RCF P50 and P80 ranges align more closely with realized outcomes than the original deterministic baselines in the illustrative cases. This evidence demonstrates that probabilistic modeling provides a more accurate representation of risk than single-point estimates.

A commonly cited weakness of traditional forecasting—identified by GAO (2020), CAPE (2022), and multiple independent assessments—is its reliance on deterministic, single-point estimates. These figures can create a false sense of precision while masking the uncertainty inherent in complex defense programs. As a result, decision-makers often anchor on optimistic baselines that are frequently breached, contributing to the cost overruns documented throughout this research. AI-RCF mitigates this tendency by presenting a range of plausible outcomes, making uncertainty explicit and helping leaders calibrate expectations more realistically.

AI-RCF replaces false precision with credible, risk-adjusted ranges that better reflect the uncertainty inherent in complex defense programs. Figure 7 reconceptualizes the forecast as a probabilistic "landing zone" for a program's final cost, and this visualization approach ensures clarity by presenting each program in a separate small-multiples panel. Each panel displays three key data points: the Initial Estimate (blue circle), representing the original optimistic baseline; the AI-RCF Forecast Band (shaded area), representing the likely range of outcomes between the P50 and P80 projections; and the actual final cost (red "X"), representing the realized outcome. Presenting the data in this format makes the divergence between deterministic baselines and probabilistic forecasts easier to interpret and highlights how AI-RCF provides earlier visibility into likely cost-growth trajectories.

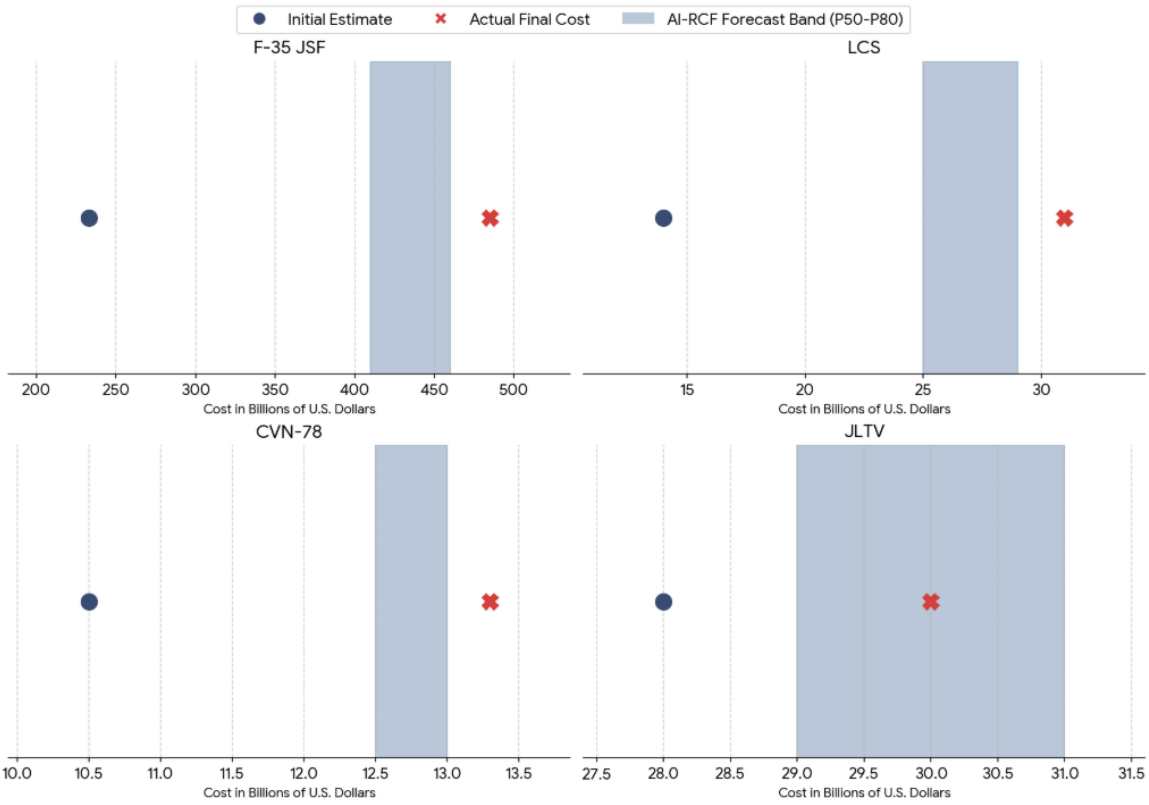

Figure 7. Comparing Initial Estimates Against AI-RCF Forecast "Landing Zones." Adapted from DoD (2023g), GAO (2015, 2020, 2025), Reeves (2025). Image generated using Microsoft Copilot (2026).

This visualization highlights the comparative advantages of a probabilistic approach. For high-risk programs like the F-35 JSF, the initial estimate was dramatically

lower—$233 billion compared to the realized $485 billion, a difference of more than 100%. In this illustrative simulation, the realized cost falls within the depicted AI-RCF's predicted forecast band, demonstrating that the model could have provided a realistic and actionable warning of the likely cost range from the outset.

Similarly, for the CVN-78 and LCS programs, the actual outcomes landed very close to the pessimistic P80 forecast, underscoring how AI-RCF captured the elevated risk associated with immature technologies, below-threshold TRLs, and concurrency. In contrast, the relatively stable JLTV program showed tight alignment among the initial estimate, the AI-RCF forecast band, and the realized cost, illustrating the model's ability to confirm low-risk profiles as effectively as it identifies high-risk ones.

By providing a credible "landing zone" rather than relying on a single deterministic point estimate, AI-RCF equips leaders to make decisions with a clear-eyed understanding of potential risk trajectories. This visual reinforces the strategic value of probabilistic forecasting and supports the case for its institutionalization within DoD acquisition planning.

To ensure trust and transparency, AI-RCF models should undergo rigorous back-testing against historical program outcomes and sensitivity analysis to identify feature importance. These validation steps strengthen model explainability and help decision-makers understand the drivers behind forecast outputs. Additionally, by automating reference-class selection, AI-RCF reduces cognitive biases such as optimism bias and anchoring, enabling more objective and empirically grounded early-phase estimates.

In addition to expenditure growth, schedule slippage remains a persistent challenge across major defense programs, often exceeding initial estimates by years. As shown in Figure 8, years of schedule slippage relative to initial program estimates. The F-35 JSF and LCS experienced the most significant delays, while the JLTV and CVN-78 showed more moderate slippage. These patterns highlight the limitations of traditional forecasting and the need for probabilistic, data-driven approaches like AI-RCF.

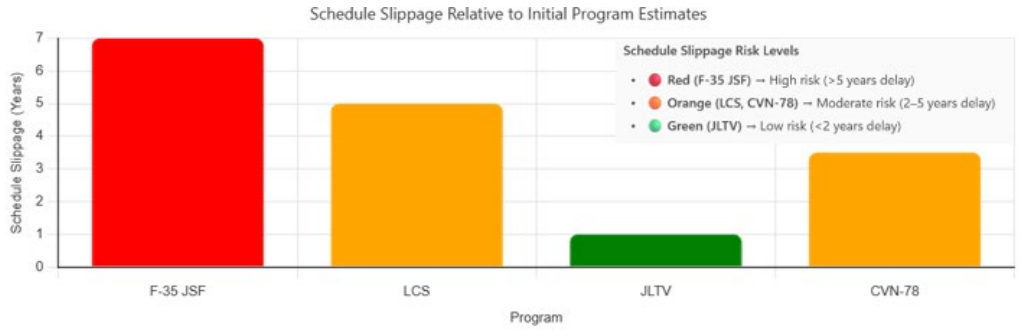

Figure 8. Schedule Slippage Relative to Initial Program Estimates. Adapted from DoD (2023g), GAO (2015, 2020, 2025), Reeves (2025). Image generated using Microsoft Copilot (2026).

These discrepancies reflect the optimism bias and planning fallacy discussed earlier in the capstone project, where programs systematically underestimate cost and schedule risk despite historical evidence to the contrary. AI-RCF mitigates these tendencies by grounding forecasts in empirical analogs and generating probabilistic ranges that make uncertainty explicit, enabling decision-makers to plan with greater realism and resilience.

The preceding analysis of the F-35, LCS, JLTV, and CVN-78 programs illustrates the specific benefits of the AI-RCF framework using simulated, in-sample data. These benefits include earlier visibility into likely cost and schedule growth, more accurate representation of uncertainty through P50 and P80 ranges, and improved alignment between forecasted and realized outcomes. While such analysis demonstrates the conceptual value of the approach, a rigorous assessment of any forecasting method requires testing its performance on data not used to train or calibrate the model.

C. OUT-OF-SAMPLE BACK-TEST VALIDATION USING MSARS

While the retrospective case studies in Section B established the baseline mechanics of the RCF framework, evaluating how the approach can be applied in an out-of-sample context is essential for assessing its broader relevance to defense acquisition. This step helps mitigate concerns related to overfitting and demonstrates how probabilistic, distribution-based methods can be extended to programs that were not part of the original reference-class construction. To support this objective, this study examines a holdout set of ten MSARs spanning aviation, shipbuilding, communications, and

modernization portfolios. These programs are intentionally excluded from reference-class development and are used solely to illustrate how the framework can be applied across a diverse acquisition environment.

1. Framing of Holdout Findings

The holdout set is used to demonstrate how probabilistic forecast ranges derived from reference-class distributions could be interpreted alongside realized unit-cost outcomes across programs with varying characteristics. Rather than serving as a formal test of predictive accuracy, this analysis illustrates how a range-based forecasting approach enables structured interpretation of cost variability in an out-of-sample setting. The emphasis is therefore placed on understanding dispersion, baseline sensitivity, and coverage concepts, rather than assessing point estimate precision.

2. Holdout Dataset and Sources

The out-of-sample set consists of acquisition programs spanning varied mission areas using MSAR unit-cost reporting (DoD, 2023a, 2023b, 2023c, 2023d, 2023e, 2023f, 2023g, 2023h, 2023i, 2023j; 2024a), B-52 Radar Modernization Program (RMP); MQ-25; MH-139A; CH-47F Block II; Ground/Air Task Oriented Radar (G/ATOR); Improved Turbine Engine Program (ITEP); VC-25B; E-2D Advanced Hawkeye (AHE); CVN-78; and Multifunctional Information Distribution System (MIDS). These programs were excluded from reference-class construction and serve solely as a holdout set for empirical evaluation. Although CVN-78 is discussed elsewhere in the capstone as an illustrative acquisition case, its inclusion here is limited to independently reported cost outcomes and does not influence reference-class formation.

3. Extraction Targets and Constructed Variables

For each MSAR, the study extracts unit-cost reporting elements required to evaluate realized cost growth in a standardized manner. The primary dependent measures are Program Acquisition Unit Cost (PAUC) and Average Procurement Unit Cost (APUC), which serve as auditable indicators of observed acquisition outcomes. The resulting percentage changes represent *ex post* realizations rather than forecasts. These values provide a basis for examining how observed outcomes relate conceptually to

probabilistic forecast ranges derived from a reference class, while preserving traceability to official acquisition reporting.

a. Out-of-sample evaluation procedure

The holdout analysis is structured to reflect how an acquisition oversight body could apply AI-enabled RCF concepts in practice. In this context, reference-class distributions and associated uplift concepts are treated as independently determined from the development dataset. The illustrative workflow involves: (a) identifying the relevant baseline unit-cost anchor (PAUC or APUC), and (b) specifying that selected confidence levels (e.g., P50 or P80) correspond to percentile-based uplifts that would be applied to the baseline to form probabilistic forecast ranges. Within this study, these steps are presented conceptually rather than implemented program-by-program. Accordingly, the analysis uses observed MSAR outcomes to examine dispersion and baseline sensitivity, rather than conducting a formal comparison against computed probabilistic forecast bands.

b. Illustrative governance relevance

Several programs in the holdout set include explicit unit-cost breach reporting, underscoring the operational relevance of governance-oriented forecasting tools. For example, the MH-139A MSAR documents a Nunn–McCurdy breach driven by quantity reductions and associated increases in PAUC and APUC. This case highlights how programmatic decisions can materially alter unit-cost outcomes and illustrates the value of incorporating probabilistic perspectives when assessing acquisition risk.

c. Results reporting and reproducibility

The extracted holdout dataset is presented in Table 6, with each entry traceable to the corresponding MSAR source. The table reports realized PAUC and APUC percentage changes across both current MSAR baselines and original Milestone B baselines. Presenting outcomes relative to multiple baselines highlights how differences in baseline selection influence the apparent magnitude of cost growth. Specifically, larger percentage changes relative to the original baseline may reflect a smaller initial denominator rather than additional post–re-baseline growth.

This approach supports transparency, reproducibility, and future analytical extension, while providing an empirical foundation for examining cost dispersion across a heterogeneous portfolio.

Table 6 Realized Percentage Changes Across Alternative Baselines.

Program	PAUC %Δ (Current Baseline)	APUC %Δ (Current Baseline)	PAUC %Δ (Original Baseline)	APUC %Δ (Original Baseline)	Magnitude Tier (Absolute %Δ, Descriptive only)
B-52 RMP	0.31%	0.41%	17.86%	20.01%	High (15–30%)
MQ-25	0.66%	9.25%	0.66%	9.25%	Moderate (5–15%)
MH-139A	22.12%	15.17%	11.72%	2.42%	High (15–30%)
CH-47F Block II	(10.11%)	(9.24%)	(10.11%)	(9.24%)	Moderate (5–15%)
G/ATOR	2.82%	3.81%	20.05%	7.01%	High (15–30%)
ITEP	6.45%	6.21%	6.45%	6.21%	Moderate (5–15%)
VC-25B	2.20%	—	2.33%	—	Low (≤5%)
E-2D AHE	4.83%	(3.88%)	13.19%	(7.23%)	Moderate (5–15%)
CVN-78	3.43%	13.47%	(13.30%)	(19.33%)	High (15–30%)
MIDS	1.50%	(0.80%)	(74.64%)	(71.73%)	Extreme (>30%)

Source: Adapted from DoD MSARs (2023–2024a). Magnitude tiers reflect the absolute size of realized percentage change and are descriptive only; they do not indicate program success, failure, or forecast accuracy.

Table 6 shows that realized outcomes span a wide range of cost-growth magnitudes, consistent with the variability expected across a cross-program acquisition portfolio. When evaluated against current MSAR baselines, many programs appear comparatively stable. However, when compared to original Milestone B baselines, several programs exhibit substantially greater dispersion. This contrast demonstrates how identical realized outcomes can appear materially different depending on the baseline used for evaluation, reinforcing the importance of consistent reference points in both cost estimation and performance assessment.

d. Implications for acquisition governance

The observed baseline sensitivity is not merely an analytical observation but a practical governance challenge. Evaluating programs against recently adjusted baselines can compress the perceived variance in cost outcomes, while comparisons to original commitments often reveal greater dispersion. Because acquisition oversight depends on credible and comparable measures of performance, reliance on shifting baselines can obscure underlying uncertainty.

In contrast, a distribution-based framework such as RCF maintains explicit representation of uncertainty through probabilistic ranges, enabling more consistent comparisons across programs and over time. The variability observed in the holdout set therefore provides a direct rationale for adopting probabilistic approaches within acquisition governance: the value lies not in producing a single "correct" estimate, but in bounding uncertainty in a structured, evidence-based manner.

e. Scope and limitations of the analysis

This study does not implement a fully trained artificial intelligence model capable of generating program-specific P50 or P80 forecasts for the holdout set. Developing such capability would require access to comprehensive historical MDAP cost data and the construction of production-grade machine-learning pipelines, which are beyond the scope of this capstone. Instead, the analysis demonstrates how empirically grounded reference-class concepts can be applied using real acquisition outcomes, providing a governance-focused proof of concept rather than a benchmark of predictive accuracy.

Across the holdout set, realized outcomes vary in both magnitude and direction, reflecting the complexity and diversity of defense acquisition programs. This variability is consistent with the central premise of RCF: while individual program trajectories remain uncertain, portfolio-level distributions can still provide meaningful decision support. From a governance perspective, the key objective is not to achieve exact point predictions, but to ensure that uncertainty is represented in a manner that reasonably bounds potential outcomes in advance. This framing preserves the distinction between probabilistic decision support and deterministic cost estimation, and supports the broader

argument for scalable, distribution-based forecasting methods within DoD acquisition processes.

D. STRATEGIC APPLICATIONS, RISKS, AND LIMITATIONS

AI–RCF extends its value well beyond improving forecast accuracy. Its ability to generate probabilistic, evidence-based insights enables broader applications across acquisition planning, from validating technology readiness to shaping contract strategies and sustainment decisions. The discussion below outlines these strategic uses before turning to the risks and limitations that shape practical implementation.

1. Strategic Applications

TRA Validation: AI-RCF estimates technology maturation likelihood using P50/ P80 confidence intervals, supporting more defensible milestone decisions and reducing reliance on subjective assessments (DoD, 2025; GAO, 2020).

Contract Strategy Optimization: Patterns in historical contract performance reveal conditions under which fixed-price or cost-type structures are more effective. AI-RCF helps acquisition teams align contract type with program risk posture, improving strategy formulation.

2. Risks and Limitations

While these applications demonstrate clear benefits, adopting AI-RCF introduces challenges that must be addressed to ensure responsible and credible use. Data quality, institutional resistance, and ethical considerations can affect implementation and trust. This section examines these risks and limitations to provide a balanced perspective on feasibility.

- Data Quality: AI models are only as reliable as the data they are trained on. Incomplete, outdated, or biased datasets can lead to inaccurate forecasts. Ensuring data integrity and representativeness is critical for effective AI-RCF integration.
- Institutional Resistance: Adoption of AI tools can face cultural and bureaucratic resistance within the DoD. Legacy systems, entrenched practices, and skepticism toward automation can hinder implementation, requiring targeted change management and training efforts.

- Gaming the System: A significant risk associated with predictive model is the potential for users to attempt to 'game the system.' As program teams become familiar with the AI-RCF model's key features (e.g., TRL, concurrency), they may be incentivized to manipulate these inputs to secure a more favorable forecast. This risk may not be eliminated entirely, but it can be mitigated through robust governance. As recommended, establishing immutable, time-stamped data inputs from authoritative sources (like formal TRAs) and maintaining strict audit trails for all forecast-related data can deter such behavior. Furthermore, the model's reliance on a wide array of features, including text-derived sentiment from qualitative reports, makes it far more difficult to manipulate than a simple checklist.
- Ethical and Policy Concerns: The use of AI in decision-making raises ethical questions around accountability, fairness, and transparency. Policies must be updated to address issues such as algorithmic bias, data privacy, and the explainability of AI-generated forecasts (Commission on PPBE Reform, 2024; DoD, 2025).

The insights from these case studies and comparative analyses demonstrate the predictive advantage of AI–RCF and its potential as a decision-support tool. Building on these findings, the following chapter translates conceptual and empirical evidence into policy recommendations and implementation strategies for institutional adoption across the DoD acquisition enterprise.

THIS PAGE INTENTIONALLY LEFT BLANK

VI. RECOMMENDATIONS

The policy recommendations in this section translate the findings from Chapters IV and V into actionable reforms that strengthen forecasting discipline, improve early-phase decision accuracy, and enhance alignment between acquisition and sustainment outcomes. Each recommendation is designed to address systemic weaknesses identified in the comparative case analysis—particularly the persistent reliance on deterministic estimates, inconsistent treatment of technical maturity, and limited integration of empirical risk evidence into milestone decisions. Together, these recommendations outline a practical pathway for institutionalizing AI-enhanced RCF within existing DoD governance structures, enabling more realistic baselines, stronger risk management, and improved life cycle affordability.

A. STRENGTHENING ACQUISITION-TO-SUSTAINMENT FORECASTING

Current DoD acquisition practice too often underweights sustainment risk at early milestones, creating a disconnect between development decisions and long-term readiness outcomes. Wong et al.'s (2022) review of three decades of acquisition reforms highlights that successful improvement requires tailoring strategies to program context, incentivizing the acquisition workforce, and engaging a broader industrial base to leverage innovation. These insights reinforce the need for institutionalizing AI-RCF as a standard analytic input, helping embed forecasting rigor and adaptability across acquisition and sustainment planning. Portfolio analyses reveal recurring cost escalation and schedule overruns, particularly in programs that entered development or rapid prototyping with technologies below maturity thresholds, conditions that appear to correlate with elevated risk. The recommendations in this chapter build directly on the Proposed AI–RCF Model Structure introduced in Chapter IV (see Figure 4). That architecture operationalizes AI-RCF through automated clustering, probabilistic modeling, and governance features, making it the technical backbone for institutionalizing data-driven forecasting across DoD acquisition (DoD, 2025; GAO, 2025).

DoD policy already establishes the framework to address this gap. DoDI 5000.91 mandates product support planning and makes the LCSP the principal document for system sustainment strategy (DoD, 2021)—emphasizing data-driven decisions, early sustainment integration, and defined responsibilities for program managers and product support managers. The DoD (2024b) Product Support Manager (PSM) guidebook further directs programs to maximize readiness at the lowest O&S cost by integrating sustainment considerations at the outset and documenting them in the LCSP.

To operationalize these policies, this capstone project recommends institutionalizing AI-RCF as a standard analytic input to LCSPs and milestone documentation. AI-RCF would (1) quantify technology-maturation and integration risk using outside-view analogs, directly supporting Milestone A/B decisions and TRA findings; (2) produce probabilistic P50/P80 bands for development and O&S expenditures to inform product support strategies; and (3) surface contract-type patterns (e.g., when fixed-price vs. cost-plus historically performs better for similar systems) to align incentives with sustainment outcomes. This aligns with GAO's TRA guidance, which associates technologies below maturity thresholds with elevated cost and schedule risk and provides the data-driven transparency DoD policy prescribes. Table 7 demonstrates how commonly observed sustainment gaps—such as weak early tradeoff analysis, immature technologies, and fragmented governance—can be systematically addressed through targeted AI–RCF interventions and integrated into LCSPs.

Table 7 Mapping Common Sustainment Gaps to Interventions.

Sustainment Gap	AI–RCF Intervention	LCSP Application
Sustainment planning not driving early tradeoffs	Generate probabilistic O&S cost bands from analogous programs	Use in LCSP sections on sustainment KPP/KSAs and reliability strategy
Immature technologies and downstream cost/schedule risk	Forecast technology maturation probability (TRL distributions) using reference classes	Link to risk register, reliability growth planning, and product support decisions
Cost growth and delayed capability across portfolio	Use reference classes to benchmark schedule realism and O&S exposure	Align LCSP schedule-to-supportability decisions (spares, data, depot activation)

Weak linkage between product support strategy and incentives	Identify incentive structures from analogous programs with strong sustainment outcomes	Strengthen LCSP Product Support Strategy and BCA justification
Fragmented governance into sustainment drivers	Establish auditable forecasting workflow using historical data and model updates	Improve LCSP metrics governance, feedback loops, and sustainment reviews

Source: Adapted from DoD (2024b) PSM guidebook; DoD (2021), GAO (2020, 2024a, 2025).

Table 7 demonstrates a consistent relationship between identified sustainment gaps, targeted AI–RCF interventions, and their corresponding LCSP applications. Across each row, persistent acquisition challenges—such as weak early tradeoff analysis, immature technologies, and fragmented governance—are directly mapped to specific probabilistic, data-driven solutions. For example, gaps related to immature technologies are addressed through forecasting technology maturation probabilities, which are operationalized through linkage to risk registers and reliability growth planning. Similarly, cost growth and schedule instability are addressed through reference class benchmarking, enabling more realistic alignment of supportability decisions such as spares provisioning and depot activation timing.

These relationships illustrate how AI–RCF extends beyond forecasting to function as a structured decision-support mechanism. By systematically linking observed gaps to probabilistic interventions and governance processes, the framework translates empirical evidence into actionable sustainment strategies. This reinforces the central argument of this capstone: embedding data-driven, reference-class-based analysis within LCSP processes strengthens alignment between early acquisition decisions and downstream sustainment outcomes.

More broadly, institutionalizing AI–RCF provides a mechanism for mitigating persistent cost growth and schedule delays identified in GAO portfolio assessments. By grounding decisions in historical program performance rather than deterministic assumptions, probabilistic forecasting enables earlier identification of cost, schedule, and supportability risks. When incorporated into LCSP processes, these methods improve alignment of product support strategies, resource allocation, and sustainment metrics with

empirically derived outcomes. Collectively, this mapping demonstrates how AI–RCF can be operationalized as a standard analytical input across both acquisition and sustainment planning.

B. PORTFOLIO-LEVEL RISK MANAGEMENT AND CONTINUOUS LEARNING

AI-RCF's utility extends beyond individual program forecasting to enterprise-level decision-making. By aggregating probabilistic forecasts across programs, senior leaders can better manage risk at the portfolio level rather than relying solely on isolated contingency reserves. This approach enables dynamic resource allocation during the PPBE process, allowing trade-offs between cost, schedule, and capability under quantified uncertainty. For example, portfolio managers can model scenarios such as funding multiple programs at P50 confidence versus fewer programs at P80, aligning investment strategies with risk tolerance and strategic priorities.

Equally important, AI-RCF can support a continuous learning loop. Each completed program feeds actual outcomes back into the reference class database, refining future forecasts and institutionalizing lessons learned. This helps shift acquisition from a reactive posture toward a more predictive, adaptive enterprise—one that systematically improves forecasting accuracy over time. By embedding this feedback mechanism into governance structures, the DoD can evolve toward a true learning organization, reducing systemic expenditure growth and accelerating capability delivery across the life cycle.

C. ALIGNMENT WITH GAO AND DOD POLICY

AI-RCF aligns with the GAO's four pillars for high-quality cost estimates. It is comprehensive because it incorporates life cycle costs and historical risk factors rather than relying solely on point estimates. It is well-documented through the generation of auditable model cards and data lineage, ensuring transparency and reproducibility. It is accurate because forecasts are anchored in empirical distributions and validated through back testing against historical program outcomes.

Embedding AI-RCF into milestone reviews reinforces the principles outlined in DoDI 5000.85, which emphasize data-driven analysis and active risk management. By

replacing deterministic point estimates with probabilistic forecasts, AI-RCF enables decision-makers to tailor acquisition strategies based on quantified risks moving beyond manual estimation toward a transparent, data-driven discipline. This approach strengthens accountability and aligns with the Department's broader push toward transparency and evidence-based decision-making.

D. FORMALIZE AI-RCF IN ACQUISITION POLICY

To close the gap between optimistic baselines and realized outcomes, the Department should amend the 5000-series to recognize AI-RCF as an approved—and for covered programs, required—analytic for cost, schedule, and risk estimation at Milestones A/B (and pathway decision points). Portfolio data show the stakes: the U.S. GAO's *Weapon Systems Annual Assessment* (2025) confirm persistent portfolio-level overruns despite repeated reform cycles and continued slippage to initial capability, trends indicating potential structural underestimation of risk during early planning. Codifying AI-RCF would institutionalize the use of probability distributions and auditable outside-view evidence as part of the acquisition record, rather than discretionary add-ons. Because DoDI 5000.91 already makes the LCSP the principal product-support document and stresses data-driven sustainment planning, inserting an AI-RCF requirement aligns with existing governance instead of creating a new gate; it simply raises the evidentiary bar at the gates we already have. In parallel, DoD (2024b) PSM guidebook frames the goal as maximizing readiness at the lowest O&S cost, which AI-RCF directly supports by quantifying sustainment-relevant drivers early.

1. Integrate AI-RCF into Key Acquisition Processes

AI-RCF should be embedded in the analytic heart of three mandatory artifacts, so decision authorities receive risk-adjusted views at the moments that matter. In TRA, programs could quantify the probability that each CTEs reaches the required TRL by a specific date, transforming TRA from a binary maturity check into a probabilistic forecast of technology risk and burn-down. That change is directly responsive to the GAO's TRA guidance, which links technologies below maturity thresholds at development start to systematic expenditure growth and delays. In Analyses of

Alternatives, alternatives would be compared using AI-RCF cost and schedule distributions (not point estimates), making the trade space explicit under uncertainty. In Test and Evaluation Master Plans, reliability-growth and test phasing would be stress-tested against AI-RCF schedule bands, so entrance/exit criteria reflect realistic timelines. Together, these changes would help replace inside-view optimism with more evidence-based assessments at each gate—an approach consistent with the GAO's findings in 2024–2025 that programs continue to enter development and rapid pathways with low maturity and then pay for it in time and money.

Other governments, including the UK and Australia, have institutionalized RCF as standard practice (ICG, 2016). The DoD can build on this precedent by leveraging AI to scale and automate the process, ensuring that forecasting realism is applied consistently across the acquisition life cycle.

2. Establish Governance and Standards

To ensure AI-RCF is reproducible, explainable, and auditable across the enterprise, the Department should mandate a minimal, enforceable data and model governance framework tied to artifacts already governed by policy. This framework should include the following components:

- Model Cards: Document model purpose, feature set, and validation, including back-testing results (e.g., mean absolute error and confidence interval accuracy). These cards provide an auditable record for oversight and reproducibility.
- Data Cards: Record data sources, lineage, and quality checks, ensuring traceability across CADE, EVAMOSC, and MSAR datasets. Each card includes coverage metrics (e.g., percentage of missing values) to flag data integrity issues.
- Explainability Tools: Use SHAP values and attention-based methods to highlight key drivers of forecasts. For example, SHAP analysis ranks TRL and contract type among the top three predictors of cost growth, enabling decision-makers to understand why forecasts vary and which factors exert the most influence.

3. Align Incentives with Forecasting Discipline

Forecasting discipline is more likely to persist when it is supported by appropriate incentives. Program leadership performance plans and award-fee structures could

reference AI-RCF bands explicitly rewarding programs that deliver within P80 expectations (or execute approved mitigations when they drift) and tightening oversight when performance falls outside established risk thresholds. Contract strategies also reflect reference-class evidence: where analogous efforts show better outcomes with performance-based arrangements that tie payout to availability and O&S outcomes, those incentives could be favored; where uncertainty is high, staggered incentive structures can be tied to meeting risk-reduction thresholds rather than calendar dates. These steps confront the pattern GAO documents—rising portfolio costs and protracted time to initial capability— embedding realistic, probabilistic expectations into accountability frameworks for personnel and contractual agreements. DoDI 5000.91 already endorses performance-based life cycle support; AI-RCF simply supplies the empirical baseline to make those incentives credible.

4. Expanding AI-RCF Utility

Once institutionalized, AI-RCF could extend beyond development forecasting to inform sustainment reviews and life cycle planning, enabling proactive adjustments to product support strategies. Integrating AI-RCF into sustainment reviews could enable earlier adjustments to product-support strategies (spares, technical data, depot activation timing) and creates a learning loop as actuals feed back into reference classes. At the portfolio level, aggregating program bands allows senior leaders to run risk-aware program objective memorandum (POM) trades, prioritizing resources toward alternatives with the strongest risk-adjusted returns, rather than the rosiest single-point claims. Because the DoD (2024b) PSM guidebook defines the objective as maximizing readiness at the lowest O&S cost and DoDI 5000.91 makes LCSP the vehicle for that objective, AI-RCF offers a quantitative mechanism that supports the intent of those policies—extending its value from development realism to end-to-end sustainment and portfolio management.

Figure 9 highlights four operational use cases that integrate probabilistic, outside-view evidence into acquisition decision-making. The diagram organizes these use cases into a continuous decision-support loop consisting of portfolio analysis, validating TRAs, contract evaluation, and milestone support, illustrating how AI–RCF outputs are

consumed across key decision points. This framework aligns with DoD 5000.91 life cycle sustainment plan governance and addresses persistent risk trends identified in recent GAO portfolio and sustainment reports.

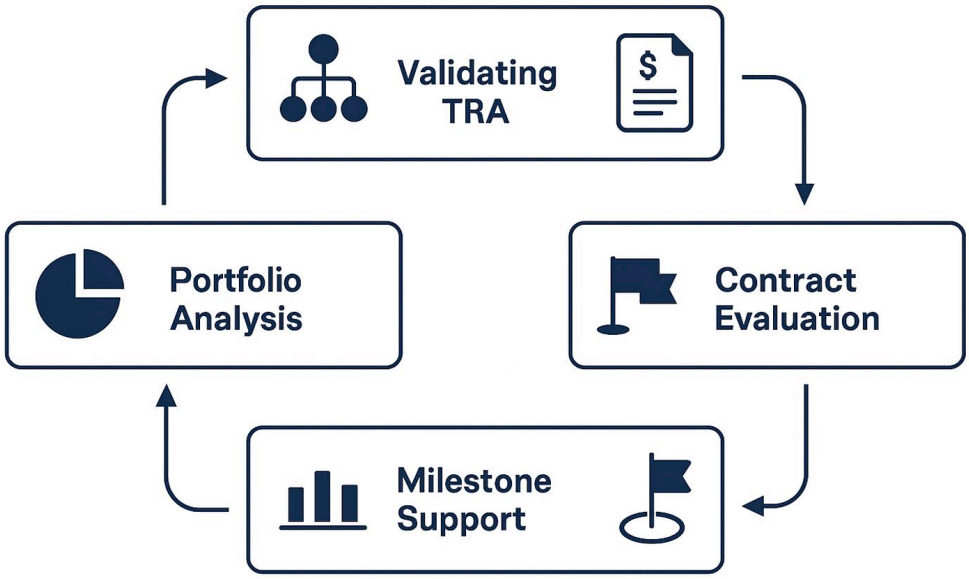

Figure 9. Operational Use Cases for Integration Across Acquisition Decision Points. Adapted from DAU (2024); DoD (2025); GAO (2020, 2025), and Flyvbjerg et al. (2016). Image generated using Microsoft Copilot (2026).

Validating TRAs quantifies the likelihood of maturing CTEs to target TRLs on time; Milestone support replaces single-point estimates with P50/P80 cost and schedule bands at A/B decisions; Contract evaluation uses reference-class evidence to select and tailor incentive structures that protect availability and O&S outcomes; and Portfolio analysis aggregates program-level bands to inform risk-aware POM trades. Together, these applications help operationalize existing policy expectations (DoD, 2021) and address portfolio shortfalls documented by the GAO's recent annual assessments and sustainment reports.

E. IMPLEMENTATION ROADMAP

To operate AI-RCF within existing DoD acquisition frameworks, a phased implementation strategy is required. The phased implementation strategy leverages the layered architecture described in Chapter IV (see Figure 4), ensuring that data governance, feature engineering, and modeling components are integrated into DoD

acquisition workflows. This alignment allows the roadmap to translate conceptual design into operational capability (DoD, 2025; GAO, 2020).

- Pilot Programs: Launch AI-RCF pilots within select acquisition portfolios (e.g., Navy aviation, Army ground vehicles, or Space Force systems) to validate methodology, refine models, and build institutional confidence.
- Training and Workforce Development: Develop tailored training modules for cost analysts, program managers, and milestone decision authorities on interpreting probabilistic forecasts and integrating AI-RCF into decision briefs.
- Infrastructure and Cybersecurity: Invest in secure, cloud-based platforms for AI-RCF deployment, ensuring compliance with DoD cybersecurity standards and enabling cross-Service data sharing.

To operate AI-RCF within existing DoD acquisition frameworks, a phased implementation strategy is required. This strategy can help ensure a deliberate, scalable rollout that builds institutional confidence and technical capability over time. Figure 10 presents this sequential, four-phase implementation roadmap. The roadmap begins with (1) Foundational Setup & Pilot Program before moving to (2) Initial Rollout & Feedback Integration. These stages prepare the enterprise for (3) Full-Scale Implementation and, ultimately, a state of (4) Continuous Improvement & Optimization. Each phase includes key actions, stakeholders, and success metrics to guide adoption.

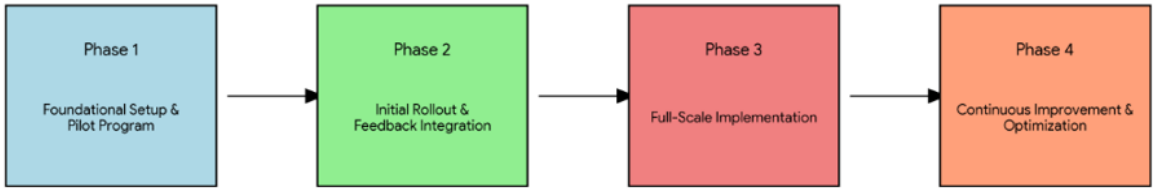

Figure 10. Phased Implementation Roadmap for AI-RCF Integration. Adapted from DAU (2024), DoD (2025), Flyvbjerg et al. (2016), and GAO (2020, 2025). Image generated using Microsoft Copilot (2026).

Phase 1 begins with pilot programs in select MDAPs to validate forecasting accuracy and user adoption. Phase 2 focuses on enhancing data infrastructure, including access to historical MSARs, DAES reports, and TRL assessments. Phase 3 aligns policy and governance structures, integrating AI-RCF into DoDI 5000.85 and CAPE guidance. Finally, Phase 4 institutionalizes AI-RCF across acquisition pathways, embedding it into milestone reviews, TRA validation, and contract strategy decisions. This roadmap

outlines a phased approach that integrates AI-RCF into existing acquisition workflows while ensuring compliance with policy and governance standards.

F. METRICS FOR SUCCESS

To evaluate the effectiveness of AI-RCF integration, the DoD must adopt a set of quantifiable performance indicators that reflect improvements in forecasting realism, risk management, and institutional accountability.

- Forecast Accuracy and Risk Visibility: Demonstrated reduction in variance between planned and actual cost/schedule outcomes. Increased use of P50/P80 confidence intervals in milestone decision documentation.
- Program Performance: Reduction in Nunn-McCurdy breaches and expenditure growth across MDAPs.

Stakeholder confidence will be measured through positive assessments from decision authorities, GAO evaluations, and congressional oversight reviews. To evaluate the institutional impact of AI–RCF integration, the DoD must adopt quantifiable performance metrics that reflect improvements in forecasting realism, risk visibility, and decision accountability. These metrics could be tracked across programs and reported during milestone reviews, POM submissions, and congressional oversight cycles. Table 8 outlines four key indicators—forecast accuracy, risk visibility, program performance, and stakeholder confidence—each paired with baseline conditions and target thresholds. These metrics provide a structured framework for assessing the potential effectiveness of AI–RCF in reducing cost and schedule overruns, enhancing transparency, and strengthening oversight across the acquisition life cycle.

Table 8 Metrics for Evaluating Integration Success.

Performance Area	Indicator	Current State	Desired State
Forecast Accuracy	Cost & schedule variance between estimate and outcome	High variance (>20%)	Reduced variance (<10–15%)
Risk Visibility	Use of probabilistic forecasts in decisions	Limited or inconsistent use	Standardized use of probabilistic forecasts

Program Performance	Nunn–McCurdy breaches and cost growth	Frequent breaches (>20%)	Reduced breach rate (<5%)
Stakeholder Confidence	External assessments (GAO, oversight bodies)	Mixed or critical evaluations	Predominantly favorable assessments

Source: Adapted from DAU (2024); DoD (2025); GAO (2020, 2025).

Table 8 presents a streamlined set of performance indicators for evaluating the effectiveness of AI–RCF integration across the acquisition life cycle. Each metric captures a shift from current baseline conditions toward improved outcomes in forecasting accuracy, risk visibility, program performance, and stakeholder confidence.

As shown, AI–RCF adoption is associated with reduced cost and schedule variance, increased use of probabilistic forecasts in decision-making, fewer Nunn–McCurdy breaches, and more favorable external assessments. Together, these indicators translate conceptual forecasting improvements into measurable outcomes, providing a practical framework for tracking progress, reinforcing accountability, and enabling continuous refinement of acquisition practices.

This alignment positions AI–RCF as a cornerstone for acquisition reform, explaining the need for future research and enterprise-level adoption. By embedding the Proposed AI–RCF Model Structure into policy, governance, and performance frameworks, these recommendations ensure that the Department moves beyond conceptual design toward full-scale institutionalization of data-driven forecasting practices (DoD, 2025; GAO, 2025). The preceding recommendations provide a roadmap for institutionalizing AI–RCF within DoD acquisition frameworks. The following conclusion synthesizes these insights, highlights the strategic implications of adopting the proposed AI–RCF Model Structure, and outlines areas for future research to ensure sustained impact.

THIS PAGE INTENTIONALLY LEFT BLANK

VII. CONCLUSION

This chapter synthesizes the findings of the research and reflects on their implications for improving forecasting accuracy, risk management, and decision-making within the DoD acquisition system. It summarizes the key insights developed across the capstone project, evaluates the strategic value of integrating AI-enhanced RCF into existing governance structures, and identifies opportunities for future research. Together, these elements reinforce the central argument that probabilistic, data-driven forecasting can strengthen acquisition outcomes and better align early decisions with long-term readiness objectives.

A. SUMMARY OF KEY INSIGHTS

This capstone project has argued that chronic forecasting breakdowns in the DoD acquisition system—manifested in budget overruns, schedule delays, and risk underestimation—may reflect systemic outcomes of flawed methodologies and entrenched cognitive biases (GAO, 2020, 2025). The evidence across this study demonstrates that probabilistic methods consistently outperform deterministic baselines.

RCF is grounded in the outside view, and it offers a demonstrated means of mitigating these failures, as evidenced by its success in the UK and other international contexts (Park, 2021). Yet, within the DoD, RCF has not been institutionalized due to cultural resistance, fragmented data environments, and the reference class problem (Wong et al., 2022).

AI offers a potentially transformative solution. Through automation of reference class identification, extraction of insights from unstructured datasets, and generation of probabilistic forecasts, AI-RCF addresses the scalability and subjectivity limitations of traditional approaches (DoD, 2025). This research has shown that AI-RCF not only improves budget and timeline realism but also extends forecasting utility into strategic domains such as TRA validation and contract strategy optimization.

B. STRATEGIC VALUE OF AI-RCF INTEGRATION

The integration of AI-RCF represents more than a technical enhancement; it is a structural reform opportunity. Wong et al.'s (2022) synthesis concludes that enduring improvements require tailored strategies and rigorous, data-driven evaluation rather than compliance-driven reforms. By embedding empirical forecasting and adaptability into acquisition planning, AI-RCF supports these principles, helping address systemic weaknesses that Wong et al.'s (2022) analysis identifies as persistent drivers of expenditure growth and schedule delays. The proposed AI–RCF model structure detailed in Chapter IV provides the technical foundation for these outcomes, ensuring that institutional adoption is supported by a scalable, transparent, and auditable architecture (DoD, 2025; GAO, 2025). By entrenching realistic insights into acquisition planning, AI-RCF mitigates optimism bias, planning fallacy, and anchoring—biases that have historically distorted forecasts and undermined accountability (ICG, 2016).

AI-RCF enables decision-makers to move from deterministic estimates toward probabilistic, risk-informed planning, improving transparency, auditability, and confidence in milestone decisions. Its adoption aligns with the DoD's broader push for data-driven decision-making under the Adaptive Acquisition Framework and supports compliance with GAO and DCAPE guidance on budget estimation (DoD, 2020, 2025; GAO, 2020).

The strategic value extends beyond cost and schedule. AI-RCF can inform contract structuring, sustainment planning, and portfolio-level resource allocation, enabling the Department to anticipate risks earlier and allocate resources more effectively (Wong et al., 2022). In doing so, AI-RCF supports not only acquisition reform but also the overarching goal of maintaining technological superiority and strategic readiness in an era of great-power competition (DoD, 2022).

C. FUTURE RESEARCH

While this study establishes the conceptual and operational feasibility of AI-RCF, further research is essential to validate and scale the model across diverse acquisition contexts. The following areas warrant further exploration:

- Validate Results and Explainability: Further stress-testing and validation of the proposed AI-RCF model across diverse acquisition portfolios to refine explainability features and build decision-maker trust.
- Independent Validation Against Baselines: Comparing AI–RCF forecasts against original WBS-based engineering estimates to assess accuracy and strengthen empirical credibility across acquisition programs.
- Forecasting Novel Systems: Exploring hybrid approaches that combine AI–RCF with subject matter expert (SME) judgment or Bayesian methods to support forecasting in cases where no valid reference class exists.
- Life cycle Transition of Forecasting Methods: Examining when it is appropriate to transition from RCF-based approaches to program-specific cost estimating techniques as data availability increases over the acquisition life cycle.
- Data Governance: Developing standards for data quality, security, and interoperability across DoD systems (DoD, 2025).
- Scalability Across Portfolios: Extending AI-RCF beyond MDAP to smaller, rapid acquisition efforts.
- Integration with Emerging Technologies: Exploring synergies between AI-RCF and digital engineering, predictive logistics, and cyber risk modeling.
- Cross-sector Benchmarking: Incorporating lessons from international defense programs and commercial high-technology sectors to expand reference classes (Wong et al., 2022).

Such research may refine the methodology, strengthen institutional adoption, and help AI-RCF evolve as a trusted decision-support tool. While these research priorities chart a path for refinement and scalability, the following closing perspective synthesizes the capstone projects argument and emphasizes the strategic imperative of institutionalizing AI–RCF within DoD acquisition frameworks.

D. CLOSING PERSPECTIVE

Lingering forecasting breakdowns in defense acquisition are not inevitable—they stem from systemic reliance on subjective methods and fragmented data practices. AI–enhanced RCF offers a structural solution, embedding empirical rigor and predictive analytics into every milestone decision. Institutionalizing the proposed AI–RCF model structure could help move this concept from theory toward operational capability, enabling transparent, auditable, and risk-informed planning across the acquisition life cycle. The methodology reinforces rigor through objective validation metrics—MAE,

RMSE, and confidence interval accuracy—providing a defensible basis for comparing AI–RCF forecasts against historical outcomes.

By aligning governance, incentives, and data standards with this architecture, the Department can shift from budget control to proactive risk management, helping translate lessons learned into institutionalized practices that strengthen strategic decision-making. In an era of rapid technological change and great-power competition, forecasting with realism and agility is not optional; it is a strategic imperative (DoD, 2022). AI–RCF offers a structured approach that may help meet that imperative, supporting acquisition reform translates into measurable improvements in readiness, accountability, and warfighter support. Institutionalizing AI–RCF represents a structural reform essential for maintaining readiness and fiscal discipline in an era of accelerating technological change.

LIST OF REFERENCES

Ahn, T., & Menichini, A. A. (2022). Optimal talent management of the acquisition workforce in response to COVID-19: Dynamic programming approach. *Defense Acquisition Research Journal, 29*(1), 50–77. https://doi.org/10.22594/dau.21-871.29.01

Anton, P. S., McKernan, M., Munson, K., Drezner, J., Newberry, S., & Levedahl, A. (2020). *Assessing Department of Defense use of data analytics and enabling data management to improve acquisition outcomes* (SYM-AM-20-050). Naval Postgraduate School. https://hdl.handle.net/10945/64753

Cantarelli, C. C., Davis, K., Pinto, J. K., & Turner, N. (2025). Reference class forecasting: Promises, problems, and a research agenda moving forward. *Production Planning & Control*. https://doi.org/10.1080/09537287.2025.2578708

Chronopoulos, I., Raftapostolos, A., & Kapetanios, G. (2024). Forecasting value-at-risk using deep neural network quantile regression. *Journal of Financial Econometrics, 22*(3), 636–669.https://doi.org/10.1093/jjfinec/nbad014

Commission on Planning, Programming, Budgeting, and Execution Reform. (2024, March 6). *Full report of the Commission on PPBE Reform.* https://ppbereform.senate.gov/wp-content/uploads/2024/03/Commission-on-PPBE-Reform_Full-Report_6-March-2024_FINAL.pdf

Davis, G. A., DePriest, T. L., Gladstone, B. G., Hildreth, L. A., & Seitz-McLeese, M. G. (2020). Evaluating and predicting contract performance using machine learning: A feasibility study. *Institute for Defense Analyses.* https://www.acq.osd.mil/asda/dpc/api/docs/evaluating%20and%20predicting%20contract_final_091720.pdf

Department of Defense. (2013). *DoD Directive 7045.14: The planning, programming, budgeting, and execution (PPBE) process* (Incorporating Change 1, August 25, 2017). https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/704514p.pdf

Department of Defense. (2015). *DoD Instruction 5000.02: Operation of the defense acquisition system* (Incorporating Change 3, August 10, 2017). https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002p.pdf

Department of Defense. (2020). *DoD instruction 5000.85: Major capability acquisition.* https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/DoDi/500085p.pdf

Department of Defense. (2021). *DoD instruction 5000.91: Product support management for the Adaptive Acquisition Framework.* https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500091p.PDF

Department of Defense. (2023a). *Modernized selected acquisition report (MSAR): B-52 Radar Modernization Program (RMP).* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/B-52_RMP_MSAR_Dec_2023.pdf

Department of Defense. (2023b). *Modernized selected acquisition report (MSAR): CH-47F Block II.* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/CH-47F_Block_II_MSAR_Dec_2023.pdf

Department of Defense. (2023c). *Modernized selected acquisition report (MSAR): CVN 78 Gerald R. Ford Class Nuclear Aircraft Carrier.*https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/CVN%2078%20MSAR%20Dec%202023.pdf

Department of Defense. (2023d). *Modernized selected acquisition report (MSAR): E-2D Advanced Hawkeye Aircraft.* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/E-2D%20AHE%20MSAR%20Dec%202023.pdf

Department of Defense. (2023e). *Modernized selected acquisition report (MSAR): Ground/Air Task Oriented Radar (G/ATOR).* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/GATOR%20MSAR%20Dec%202023.pdf

Department of Defense. (2023f). *Modernized selected acquisition report (MSAR): Improved Turbine Engine Program (ITEP).* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/ITEP_MSAR_Dec_2023.pdf

Department of Defense. (2023g). *Modernized selected acquisition report (MSAR): Joint Light Tactical Vehicle (JLTV).* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/JLTV_MSAR_Dec_2023.pdf

Department of Defense. (2023h). *Modernized selected acquisition report (MSAR): Multifunctional Information Distribution System (MIDS).* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/MIDS%20MSAR%20Dec%202023.pdf

Department of Defense. (2023i). *Modernized selected acquisition report (MSAR): MQ-25 Stingray.* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/MQ-25%20MSAR%20Dec%202023.pdf

Department of Defense. (2023j). *Modernized selected acquisition report (MSAR): VC-25B.* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/FY_2023_SARS/VC-25B_MSAR_Dec_2023_v3_Cleared.pdf

Department of Defense. (2024a). *Modernized selected acquisition report (MSAR): MH-139 Grey Wolf (MH-139A).* https://www.esd.whs.mil/Portals/54/Documents/FOID/Reading%20Room/Selected_Acquisition_Reports/June_2024_MSARs/MH-139A_MSAR_30_Jun_2024_DOPSR_Edited.pdf

Department of Defense. (2024b). *Product support manager (PSM) guidebook.* https://aaf.dau.edu/storage/2024/08/Product-Support-Manager-PSM-Guidebook.pdf

Department of Defense. (2025). *DoD PPBE reform implementation plan.* https://media.defense.gov/2025/Jan/17/2003629812/-1/-1/1/DoD-PPBE-REFORM-IMPLEMENTATION-PLAN.pdf

Department of Defense, Cost Assessment and Program Evaluation. (2022). DoD Cost Estimating Guide v2.0. https://www.cape.osd.mil/files/otherGuides/DoD_CEGuidev2_FINAL_PR.pdf

Department of Defense, Cost Assessment and Program Evaluation. (2025). *Operating and Support (O&S) Cost Estimating Guide.* https://www.dau.edu/sites/default/files/2025-02/2025%20OS%20Cost%20Estimating%20Guide.pdf

Defense Acquisition University. (2024). *Use of artificial intelligence in acquisition* [Presentation]. https://www.dau.edu/sites/default/files/2024-08/Use%20of%20AI%20in%20Acq.pdf

Defense Innovation Board. (2019). *AI principles: Recommendations on the ethical use of artificial intelligence by the Department of Defense*. [White paper]. https://www.aiaa.org/wp-content/uploads/2024/12/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.pdf

Dwyer, M., Tidwell, B., Blivas, A., & Hunter, A. (2020, April). Cycle-times and cycles of acquisition reform (SYM-AM-20-046). *Proceedings of the Seventeenth Annual Acquisition Research Symposium*, Naval Postgraduate School. https://hdl.handle.net/10945/64750

Fitzenberger, B., & Wilke, R. A. (2015). Quantile regression methods. In R. Scott & S. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences.* Wiley. https://research.cbs.dk/en/publications/quantile-regression-methods/

Flyvbjerg, B. (2008). Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, *16*(1), 3–21. https://doi.org/10.1080/09654310701747936

Flyvbjerg, B. (2021). Top ten behavioral biases in project management: An overview. *Project Management Journal, 52*(6), 531–546. https://doi.org/10.1177/87569728211049046

Flyvbjerg, B., Hon, C.-K., & Fok, W. H. (2016). Reference class forecasting for Hong Kong's major roadworks projects. *Proceedings of the Institution of Civil Engineers – Civil Engineering, 169*(6), 17–24. https://doi.org/10.1680/jcien.15.00075

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. https://www.jstor.org/stable/2699986

Government Accountability Office. (2009). *Defense acquisitions: Assessments of selected weapon programs* (GAO-09-326SP).https://www.gao.gov/products/gao-09-326sp

Government Accountability Office. (2015). *Ford class aircraft carrier: Poor outcomes are the predictable consequences of the prevalent acquisition culture* (GAO-16-84T). https://www.gao.gov/assets/gao-16-84t.pdf

Government Accountability Office. (2016, March). Defense acquisitions: Assessments of major weapon programs (GAO-16-329SP). https://www.gao.gov/products/gao-16-329sp

Government Accountability Office. (2020). *Technology readiness assessment guide: Best practices for evaluating the readiness of technology for use in acquisition programs and projects* (GAO-20-195G). https://www.gao.gov/products/gao-20-195g

Government Accountability Office. (2022, Feb 24). Littoral Combat Ship: Actions Needed to Address Significant Operational Challenges and Implement Planned Sustainment Approach (GAO-22-105387). https://www.gao.gov/products/gao-22-105387

Government Accountability Office. (2023). *Weapon systems annual assessment: Programs are not consistently implementing practices that can help accelerate acquisition* (GAO-23-106059). https://www.gao.gov/assets/gao-23-106059.pdf

Government Accountability Office. (2024a). *Weapon systems annual assessment* (GAO-24-106831). https://www.gao.gov/assets/gao-24-106831.pdf

Government Accountability Office. (2024b, May 16). *The F-35 will now exceed $2 trillion as the military plans to fly it less.* GAO WatchBlog. https://www.gao.gov/blog/f-35-will-now-exceed-2-trillion-military-plans-fly-it-less

Government Accountability Office. (2025). *Weapon systems annual assessment: DoD leaders should ensure that newer programs are structured for speed and innovation* (GAO-25-107569). https://www.gao.gov/assets/gao-25-107569.pdf

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences, 622,* 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Independent Technical Risk Assessments for Major Defense Acquisition Programs, 10 U.S.C. § 4727. (formerly §2448b) (recodified 2023). https://uscode.house.gov/view.xhtml?req=granuleid:USC-prelim-title10-section4272&num=0&edition=prelim

Internal Consulting Group. (2016). *Reference class forecasting (CCS 010): Certified case study* [Presentation]. https://internalconsulting.com/store/l34pkf/ICG-CCS-010-Reference_Class_Forecasting.pdf

Jimenez, C. A., White, E. D., Brown, G. E., Ritschel, J. D., & Lucas, B. M. (2016). Using pre-Milestone B data to predict schedule duration for defense acquisition programs. *Journal of Cost Analysis and Parametrics, 9*(2), 112–126. https://www.researchgate.net/publication/306048612_Using_Pre-Milestone_B_Data_to_Predict_Schedule_Duration_for_Defense_Acquisition_Programs

Jones, S. J. (2022). *Analysis of cost and schedule estimation trends for major defense acquisition programs* (AFIT-ENV-MS-22-M-215). Air Force Institute of Technology. https://apps.dtic.mil/sti/trecms/pdf/AD1173774.pdf

Jones, S., White, E. D., Ritschel, J. D., & Valentine, S. M. (2024). Schedule and cost estimations through the decades: Are they improving? *Journal of Cost Analysis and Parametrics, 11*(1), 57–69. https://www.iceaaonline.com/wp-content/uploads/2024/11/JCAPv11i1-ScheduleCostEstimationsDecades-Jones.pdf

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). Pearson. https://web.stanford.edu/~jurafsky/slp3/ed3book_jan26.pdf

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47*(2), 263–291. https://doi.org/10.2307/1914185

Khan, L., Elshennawy, A., Furterer, D., & Cudney, E. (2024). Machine learning in aerospace and defense industries: A systematic review. *Journal of Management and Engineering Integration, 17*(2). https://www.researchgate.net/publication/385439156_Machine_Learning_ML_in_aerospace_and_defense_AD_industries_a_systematic_literature_review

Lovallo, D., & Kahneman, D. (2003). Delusions of success: How optimism undermines business decisions. *Harvard Business Review*, *81*(7), 56–63. https://hbr.org/2003/07/delusions-of-success-how-optimism-undermines-executives-decisions

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NIPS* 30). https://papers.nips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). University of California Press. http://projecteuclid.org/euclid.bsmsp/1200512992

McNicol, D. L. (2022). Can we explain cost growth in major defense acquisition programs? *Defense Acquisition Research Journal*, 29(1), 2–20. https://doi.org/10.22594/dau.21-867.29.01

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT)*. https://dl.acm.org/doi/10.1145/3287560.3287596

Natarajan, A. (2022). Reference class forecasting and machine learning for improved offshore oil and gas megaproject planning: Methods and application. *Project Management Journal, 53*(5), 1–14. https://doi.org/10.1177/87569728211045889

Office of the Under Secretary of Defense for Research and Engineering. (2025, February). Technology readiness assessment (TRA) guide. U.S. Department of Defense. https://www.cto.mil/wp-content/uploads/2025/03/TRA-Guide-Feb2025.v2-Cleared.pdf

Park, J. E. (2021). Curbing cost overruns in infrastructure investment: Has reference class forecasting delivered its promised success? *European Journal of Transport and Infrastructure Research, 21*(2), 120–136. https://doi.org/10.18757/ejtir.2021.21.2.5504

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann. ISBN: 9781558604797.

Porter, G., Gladstone, B., Gordon, C. V., Karvonides, N., Kneece, R. R., Jr., Mandelbaum, J., & O'Neil, W. D. (2009). The major causes of cost growth in defense acquisition (Vol. II: Main body) (IDA Paper P-4531). *Institute for Defense Analyses*. https://apps.dtic.mil/sti/pdfs/ADA519884.pdf

Prater, J., Kirytopoulos, K., & Ma, T. (2017). Optimism bias within the project management context: A systematic quantitative literature review. *International Journal of Managing Projects in Business, 10*(2), 370–385. https://doi.org/10.1108/IJMPB-07-2016-0063

Reeves, T. (2025, September 3). Just in: F-35 program still plagued by cost overruns, delivery delays, GAO says. *National Defense Magazine.* https://www.nationaldefensemagazine.org/articles/2025/9/3/just-in-f-35-program-plagued-by-cost-delivery-overruns-gao-says

Rubinstein, R. Y., & Kroese, D. P. (2016). *Simulation and the Monte Carlo Method* (3rd ed.). Wiley. ISBN: 978–1-118-63216-1.

Schmidt, E. (2018, April 17). *Statement of Dr. Eric Schmidt before the House Armed Services Committee*. U.S. House of Representatives. https://es.ndu.edu/Portals/75/Documents/HHRG-115-AS00-Wstate-SchmidtE-20180417.pdf

SEBoK Editorial Board. (2024). Cost estimating and analysis in systems engineering. *Systems Engineering Body of Knowledge*. https://sebokwiki.org/wiki/Cost_Estimating_and_Analysis_in_Systems_Engineering

Shamim, M. M. I., Hamid, A. B. A., Nyamasvisva, T. E., & Rafi, N. S. B. (2025). Advancement of artificial intelligence in cost estimation for project management success: A systematic review. *Modelling, 6*(2), 35. https://doi.org/10.3390/modelling6020035

Su, J., Jiang, C., Jin, X., Qiao, Y., Xiao, T., Ma, H., Wei, R., Jing, Z., Xu, J., & Lin, J. (2024). Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. *ArXiv.* https://arxiv.org/html/2402.10350v1

Valerdi, R. (2010). Heuristics for systems engineering cost estimation. *IEEE Systems Journal, 5*(1), 91–98. https://www.researchgate.net/publication/224169403_Heuristics_for_Systems_Engineering_Cost_Estimation

Välilä, T. (2024). Forecast errors and welfare conclusions based on the Flyvbjerg database. *Journal of Benefit-Cost Analysis, 15*(3), 440–455. https://doi.org/10.1017/bca.2024.29

Welch, C. (2025, August 1). *Army consolidates dozens of Palantir software contracts into one deal worth up to $10 billion.* Breaking Defense. https://breakingdefense.com/2025/08/army-consolidates-dozens-of-palantir-software-contracts-into-one-deal-worth-up-to-10-billion/

Womer, N. K., Al-Abedalla, B., Li, H., & Camm, J. (2025). Estimating the effects of ramp-up and learning from cost performance data. *Journal of Defense Analytics and Logistics,* 9(2), 114–133. https://doi.org/10.1108/JDAL-02-2025-0003

Wong, J. P., Younossi, O., Kistler LaCoste, C., Anton, P. S., Vick, A. J., Weichenberg, G., & Whitmore, T. C. (2022, Jun 16). Improving defense acquisition: Insights from three decades of RAND research (Research Report No. RRA1670-1). RAND Corporation. https://www.rand.org/pubs/research_reports/RRA1670-1.html

Zani, D., & Adey, B. T. (2025). Empirical comparison of weighted and non-weighted reference class forecasting. *Infrastructure Asset Management,* 12(3), 147–158. https://doi.org/10.1680/jinam.24.00024

Acquisition Research Program
Department of Acquisition, Finance and Manpower
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net