Use Case Evaluations for AI Adoption in the Department of War (DOW): Humans vs. Large Language Models (LLMs)

Muhammad F. Islam, Matt Yetto, Carol Pomales, Peter Schwartz
MITRE Corporation, 7525 Colshire Drive, McLean, VA, USA 22102-7539

ABSTRACT

As Artificial Intelligence Enabled Systems (AIES) become more integrated into the modern workforce, government organizations are also integrating AI capabilities into their workflows. The Department of War (DOW) will require a well-structured and efficient evaluation process to identify and prioritize their use of AI models. For the most efficient application of AI to DOW systems, the evaluation process should start with the examination of candidate AI use cases articulating mission impact and technical risk to inform down selection and investment decisions. Traditionally, use case evaluations are conducted by human subject-matter experts (SMEs), but this evaluation process can be labor and time intensive and variation in the SME's evaluation can potentially lead to unintended bias. Recent advances and availability of large language models (LLMs) present an opportunity to augment the expert evaluation process to improve speed and repeatability, but LLMs pose their own challenges and their alignment with human assessments in the DOW context has not been examined. This paper conducts an empirical study to compare human experts and LLM-augmented evaluations using a curated set of use cases for AI adoption in DOW. The team provided human SMEs and prompted LLMs with specific guidance on rating each use case on their implementation feasibility and mission impact criteria. The results of the human and LLM-assisted applications were analyzed to study evaluation patterns between the two groups and outliers. Results show that although LLM based evaluation scores differ from human expert evaluators, the LLM produces the same overall prioritization order of the AI use cases as human evaluators. A weak but statistically significant positive association is also observed between human and LLM based evaluations. Findings of this study offer insights into the potential uses and limitations of LLMs for structured evaluation tasks and help with defining the best practices for integrating AI-assisted evaluations into DOW pipelines seeking to frame AI use cases in effective and repeatable manner.

1. INTRODUCTION

Rapid growth in Artificial Intelligence (AI) has led to a new capability in numerous fields. Within the Department of War (DOW), many new Large Language Model (LLM) tools have been implemented into systems [1]. However, with how rapidly AI technology has sprouted and grown more prevalent, implementations and uses of these new tools have lagged in adoption across organizations [2], [3]. To aid in rapid investment into these new AI technologies, this study examines specific use cases within an organization and explores the utilization of AI capabilities to automate and streamline the process for AI technology investment prioritization.

Traditionally, use case prioritization is a time intensive process [4], [5]. Prioritization relies heavily on the knowledge of subject matter experts to assess mission impact and feasibility, manually scoring categories for individual use cases. While this approach can benefit from an expert's judgement and contextual understanding related to a specific domain of knowledge, it is inherently time-consuming and is difficult to scale [6]. Also, variability in expert's experience, interpretation of evaluation criteria, and potential personal biases can influence analysis to add variance to outcomes.

LLMs have recently emerged as a powerful tool that can be used to augment this process [7]. With an LLM's ability to analyze natural language descriptions, apply structured guidance and produce consistent outputs, it can be used to enhance the speed with which a use case can be examined [8]. Using LLMs to analyze the mission impact and implementation feasibility of a use case can solve some issues with the current process, but it may potentially add new difficulties to overcome. LLMs would be able to solve the scalability issue that expert analysis suffers from by ingesting, analyzing, and outputting responses quickly and automatically for many use cases. They also can remove any potential personal biases introduced by the expert analysis, as well as remove variance that can occur by analyzing inter-rater reliability. LLMs have the potential to introduce their own risks and limitations, where they may have limited contextual understanding, are sensitive to prompt design, and may not have sufficient mission context for analysis. These

limitations were motivations to develop an exploratory process to evaluate the tradeoffs of this capability based on LLM and refined using a retrieval-augmented generation (RAG) pipeline trained with test and evaluation best practices [9].

This research aims to provide a framework for using RAG-LLMs to automate the process of use case prioritization, for potential use across many different domains and contexts. The results for this automated prioritization process are compared to output derived from experts' manual prioritization to determine efficacy. For prioritization, use cases are examined under two primary lenses: mission impact and implementation feasibility. Within these lenses, thirteen separate categories measure the magnitude of operational mission impact that a successful deployment of the use case would have, and the potential technical and organizational difficulty associated with its implementation. Prioritization occurs by individually scoring categories for each use case and then comparing results across the set of use cases.

2. BACKGROUND

This research builds off previous work conducted within MITRE developing the prioritization process by determining best practices for automating it with AI. The prioritization process relies on specific use case analysis across separate categories. These categories are divided into two main topic areas - implementation feasibility and mission impact as described in Table 1.

Table 1. Description of Use Case Evaluation Categories

Category	Description
Implementation Feasibility	
Data Availability	a measure of the discoverability, accessibility, quality, and useability of data for AI to use to complete this use case.
AI Model Availability	a measure of access an organization has to various LLMs and AI models to implement solutions for use cases.
Team AI Technical Expertise	a measure of the technical capability and experience of the team tasked with implementing the AI capability for the use case.
Technology Maturity	a measure of the complexity of the system to implement the AI solution on, including amount and variety of data streams, compute resources, scope, scale, and external system dependencies.
System Complexity	a measure of the complexity of the system, including number and types of data streams, users, inputs, and outputs.
User Training Requirements	amount of training required for users to be able to use the AI implementation for the use case.
Acquisition, T&E, Approval	a measure of complexity for acquisition, T&E, and approval for use of the AI capability in this use case.
Mission Impact	
Relative Mission Impact within the DOW Mission	overall impact that a successful implementation of this use case would have on the broader organizational mission.
Improvement in Task Performance	how much investment in AI capabilities will affect the user's ability to complete the use case tasks.
Mission-to-task Dependency	the level to which the user's success or failure depends upon the capability this use case enables.
Vulnerability/Risks Introduced	the degree to which using AI capabilities with this use case could cause additional risks and vulnerabilities to completion of the use case.
Issues with Policy, Data Drift, and Uncertainty	the degree to which using AI capabilities with this use case can add issues with policy, data drift, and uncertainty to the completion of the use case.
User Experience	the degree to which users will trust outputs from AI implementation for this use case.

Use case prioritization then follows through an application of the principle of implementing the most feasible solutions first to aid in rapid adoption and improve overall organization familiarity with AI solutions before tackling investment into the more complicated use cases.

3. METHODOLOGY

This research developed a RAG-LLM evaluation tool that can help automate the AI use case analysis framework at scale. Its accuracy was measured by comparing its evaluation scores against a human evaluation baseline across a curated set of six AI use cases as part of the proof of concept's RAG data. The set of six spanned organization specific, and more general tool-based use cases. The use cases were standardized in format to ensure equivalent information exposure across evaluators. Information included with the use case was centered around pertinent stakeholders, potential impact, and potential risks or vulnerabilities associated with use case.

Human Evaluation Baseline

Pilot testing first began with a manual analysis to compare to the AI generated scoring later. In each trial, both quantitative scores and qualitative scores were collected. For each category within mission impact and implementation feasibility, written guidelines were provided to describe what score the category should receive based on an ordinal scale from 1 to 5. A score of 1 indicates the lowest mission impact or lowest implementation feasibility; a score of 5 indicates the highest mission impact or highest implementation feasibility. Between 2 to 5 evaluators with domain expertise in AI capabilities and test and evaluation received this defined ordinal scale with descriptive anchors to reduce ambiguity in results. In future work, more evaluators are needed to verify this pilot analysis, but this smaller dataset for human analysis was used as a proof of concept for the process. Definitions and examples were provided to evaluators to promote consistent interpretation. This formed the basis for the quantitative results. Qualitative results consisted of notes and assumptions from the evaluator describing what they knew and assumed about the use case and parent organization based on provided material, as it pertained to the category. With these steps taken, some variance remained across individual evaluators. For later analysis and comparison to the RAG-LLM output, individual human evaluations were averaged together. To ensure independent response, evaluators were instructed to not coordinate with one another. This average evaluator score served as the human baseline for later comparison with the RAG-LLM output.

LLM Based Evaluation Model Development

LLM integration focused on prompt design, structured guidance, and controlled response formats rather than traditional model retraining. Detailed instructions like the human-evaluator instructions were given to OpenAI gpt-oss-120b, a 120B-class open-weight reasoning LLM published by OpenAI for guidance. A RAG pipeline based on ChromaDB was implemented to provide the LLM with subject-matter expert generated evaluation best practices [10], [11], [12]. To assess the performance of the RAG pipeline, an evaluation was conducted using the 13 rubric categories shown in Table 1 as category-level queries. A ground-truth relevance file was constructed by dividing a test and evaluation best practices document into 10 sections. Each rubric category was linked to the best matched sections assigning relevance scores. The overall retrieval component achieved Recall@1 = 0.365, Recall@3 = 0.718 and Recall@5 = 0.808, where Recall@k measures the percentage of queries for which at least one relevant document appears in the top k retrieved results [13]. The mean reciprocal rank (MRR) which evaluates the average rank position of the first relevant source in the retrieved result was 0.846, and the normalized discounted cumulative gain at rank 5 (nDCG@5) which considers both relevance and ordering of the retrieved results was calculated as 0.740 [13]. These evaluations indicate that the RAG pipeline was generally able to retrieve the relevant sources within the top-ranked results. Since the source corpse for the RAG was relatively smaller, the performance varied across categories and some categories will benefit from further refinement.

The evaluation rubric varied from the human-evaluator rubric by replacing a number-based system with simple words to refer to the same value. For categories within mission impact, a score of 1 was instead prompted as very-low mission impact, 2 as low mission impact, 3 as medium mission impact, 4 as high mission impact and 5 as very-high mission impact. For categories within implementation feasibility, a score of 1 was prompted as very-low implementation feasibility, 2 as low implementation feasibility, 3 as medium implementation feasibility, 4 as high implementation feasibility, and 5 as very-high implementation feasibility. This change was needed as initial testing determined that although the qualitative output from the LLM would provide accurate reasoning for a decision, it would not produce an accurate quantitative numeric score to align with the reasoning it returned. This issue was resolved when asking the LLM for word representations of the numbers instead.

The RAG-LLM was designed to return outputs in JavaScript Object Notation (JSON) so that the extracted data files return constant output format across trials [14]. Once prompted with a use case, the LLM would iterate through the

scoring categories 25 times each. The scores for each category were then averaged across the 25 trials to produce the RAG-LLM's score for that category. This process was completed fifteen times, where each set of trials varied the temperature setting of the LLM, and the evaluation stance prompting. The LLM temperature setting is a parameter that controls the randomness and predictability of the output, ranging from 0.0 to 2.0 with a 0.5 interval. A temperature of 0.0 returns more deterministic output and 2.0 returns more varied output [8], [16]. The stance parameter determines the likelihood of the LLM's handling of evidence and uncertainties. The three stances used are Conservative, Neutral, and Optimistic. The conservative setting adheres more to the provided information; the optimistic setting is more likely to focus on potential benefits, but can potentially be more prone to hallucination as well [10]. The range of responses across temperature and stance settings can then be used to compare to human responses, to determine which setting most accurately replicated human variance. The prompts to the LLM included material to align the outputs to the domain context. The prompts included structural information, attributable data, and organizational assumptions to steer output to be relevant for the use cases' organization. For this testing, this input was specific to DOW but can be changed to fit the use cases of other organizations and domains.

Each of the six use cases were evaluated by the RAG-LLM tool 25 times with each temperature setting and evaluation stance. Temperature ranged from 0.0 to 2.0 with an interval of 0.5. Evaluations' prompt stances included conservative, neutral, and optimistic stances. In some trials, the RAG-LLM was unable to generate a response and produced errors. These trials were removed from the dataset before analysis. This phenomenon was more common in trials with higher temperature settings.

A calibration phase was first conducted on the six use cases to align scoring interpretations and reduce systematic bias. Calibration included reviewing representative evaluations, refining rubric language, and adjusting prompt phrasing where needed. Calibration ensured that observed differences between human and RAG-LLM assessments were attributable to evaluation behavior rather than misunderstanding of scoring criteria. This step also improved the reliability of cross-method comparisons. Figure 1 below illustrates a step-by-step workflow representation of this RAG-LLM use case evaluation methodology.

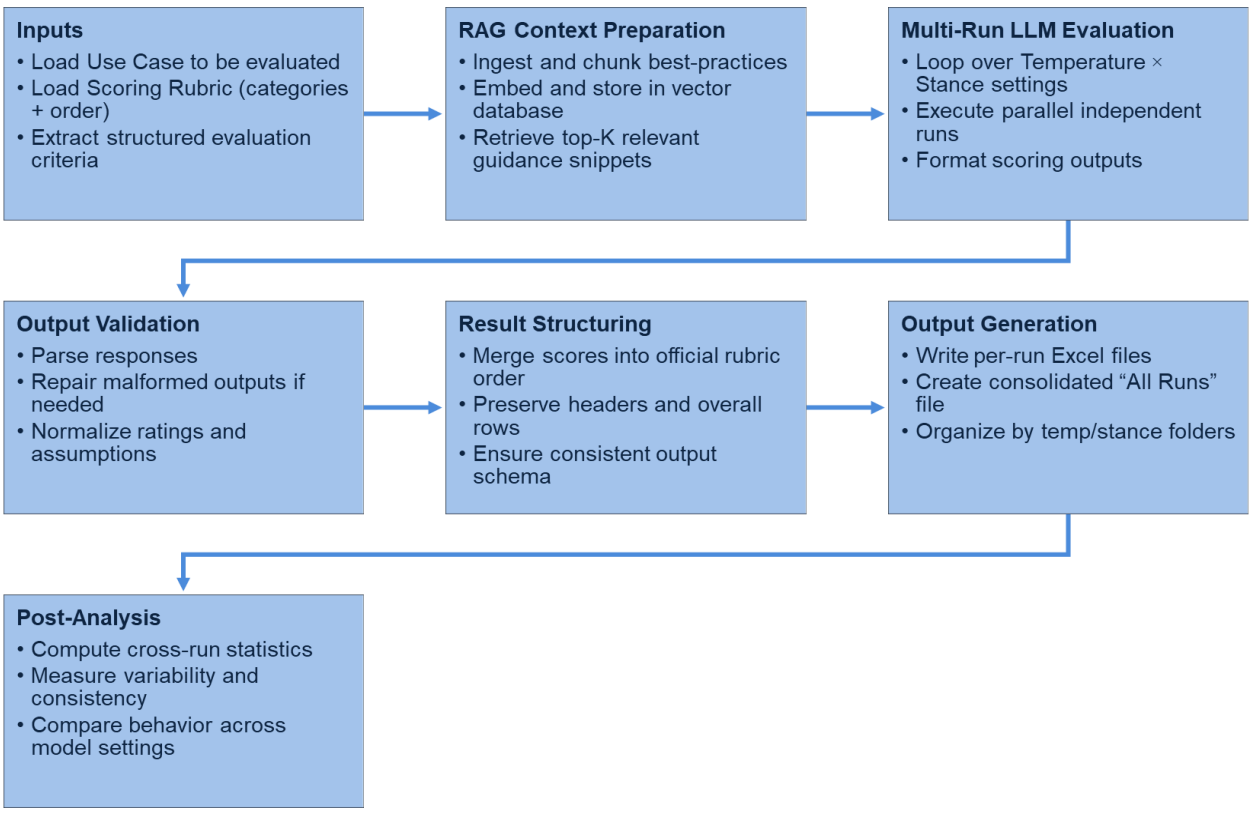


Figure 1. A Workflow Representation of the Use Case Evaluation Methodology

Analysis Metrics

The RAG-LLM based results were compared to results from human evaluation. This comparison forms the basis for an accuracy metric for the RAG- LLM's output. Additional data considered was the amount of time taken by human evaluators to complete the task compared to the RAG-LLM's evaluation time to generate a completion time metric. Accuracy is needed for the output to be valuable, but faster completion time is valuable for increasing scalability for analysis of many use cases.

4. RESULTS

Overview of Use Case Evaluations

The focus of this exploratory analysis is to evaluate the consistency and variability of evaluation ratings across the six use cases, rather than the performance of any individual use case. Each use case was evaluated using the same rubric and instructions, consisting of seven criteria of implementation feasibility and six criteria of mission impact.

Analysis of Results

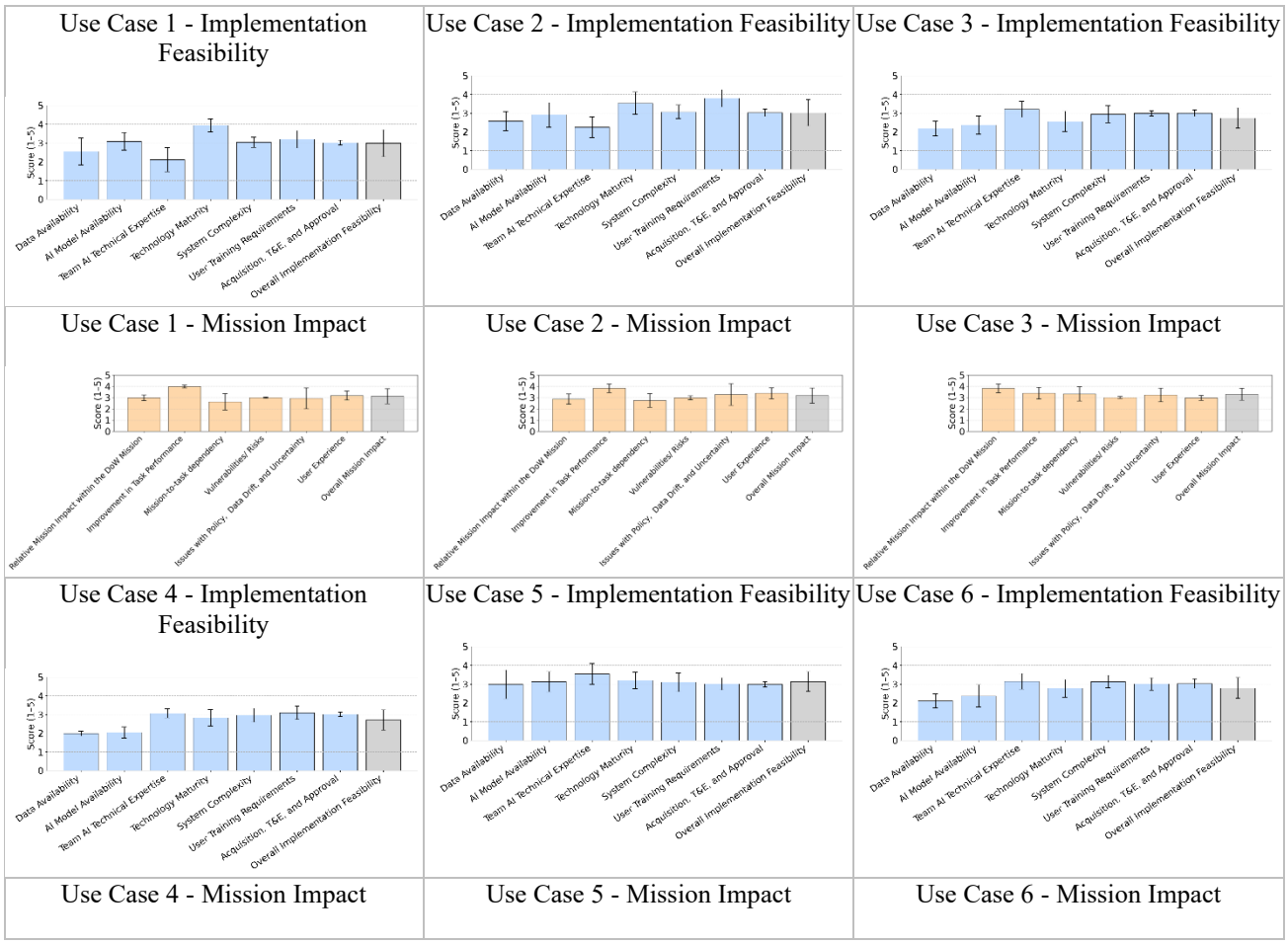
The tables below give the combined results from the use case evaluation scores of mean and standard deviations for use cases across all temperature and evaluation stance combinations. Table 2 lists the average RAG-LLM scores and their standard deviations for all categories across the 6 use cases.

Table 2. RAG-LLM Mean Scores and Standard Deviations Across Use Cases

Category	Use Case 1 (Mean ± STD)	Use Case 2 (Mean ± STD)	Use Case 3 (Mean ± STD)	Use Case 4 (Mean ± STD)	Use Case 5 (Mean ± STD)	Use Case 6 (Mean ± STD)
Implementation Feasibility						
Data Availability	2.55 ± 0.72	2.58 ± 0.52	2.17 ± 0.39	1.98 ± 0.13	2.98 ± 0.77	2.10 ± 0.37
AI Model Availability	3.09 ± 0.47	2.91 ± 0.66	2.36 ± 0.49	2.04 ± 0.29	3.13 ± 0.54	2.38 ± 0.59
Team AI Technical Expertise	2.11 ± 0.64	2.25 ± 0.55	3.21 ± 0.43	3.06 ± 0.25	3.55 ± 0.55	3.14 ± 0.42
Technology Maturity	3.94 ± 0.35	3.54 ± 0.59	2.56 ± 0.55	2.82 ± 0.43	3.20 ± 0.44	2.78 ± 0.47
System Complexity	3.04 ± 0.27	3.08 ± 0.36	2.94 ± 0.45	2.97 ± 0.37	3.09 ± 0.49	3.13 ± 0.34
User Training Requirements	3.20 ± 0.46	3.8 ± 0.46	2.99 ± 0.14	3.09 ± 0.35	3.02 ± 0.32	3.01 ± 0.32
Acquisition, T&E, and Approval	3.02 ± 0.13	3.03 ± 0.19	3.01 ± 0.17	3.01 ± 0.12	2.99 ± 0.15	3.03 ± 0.25
Overall Implementation Feasibility	2.99 ± 0.71	3.02 ± 0.70	2.75 ± 0.54	2.71 ± 0.54	3.14 ± 0.53	2.79 ± 0.55
Mission Impact						
Relative Mission Impact within the DOW Mission	2.98 ± 0.25	2.87 ± 0.46	3.82 ± 0.39	3.81 ± 0.40	3.46 ± 0.50	3.45 ± 0.50
Improvement in Task Performance	3.98 ± 0.13	3.82 ± 0.38	3.39 ± 0.51	3.39 ± 0.49	3.81 ± 0.40	3.57 ± 0.50
Mission-to-task Dependency	2.62 ± 0.72	2.74 ± 0.60	3.33 ± 0.63	2.96 ± 0.43	3.45 ± 0.60	2.80 ± 0.75
Vulnerabilities/Risks	3.01 ± 0.05	2.98 ± 0.18	3.01 ± 0.09	3.01 ± 0.21	3.10 ± 0.34	2.98 ± 0.23

Category	Use Case 1 (Mean ± STD)	Use Case 2 (Mean ± STD)	Use Case 3 (Mean ± STD)	Use Case 4 (Mean ± STD)	Use Case 5 (Mean ± STD)	Use Case 6 (Mean ± STD)
Issues with Policy, Data Drift, and Uncertainty	2.93 ± 0.93	3.27 ± 0.96	3.24 ± 0.6	3.29 ± 0.64	3.45 ± 0.77	2.95 ± 0.51
User Experience	3.18 ± 0.39	3.38 ± 0.49	2.97 ± 0.22	3.11 ± 0.31	3.32 ± 0.5	3.12 ± 0.35
Overall Mission Impact	3.12 ± 0.67	3.18 ± 0.67	3.29 ± 0.53	3.26 ± 0.52	3.43 ± 0.58	3.15 ± 0.57

Across all six use cases, mean scores for overall implementation feasibility range from 2.71 for use case 4 to 3.14 for use case 5. Average scores across all categories spread across 0.43 points on the 5-point scale. Use cases 1 and 2 are clustered near the middle point of scoring, with averages of 2.99 and 3.02. Use cases 3, 4, and 6 are scored below the middle point, with averages of 2.75, 2.71, and 2.79. Use cases 3, 4, and 6 may face comparatively higher implementation difficulties, especially in model training data availability and AI model availability categories, as these are scored lower, with averages of 1.98 and 2.38, which are lower than use cases 1, 2, and 5. Figure 2 below provides bar charts for implementation feasibility and mission impact means and standard deviations across six use cases.



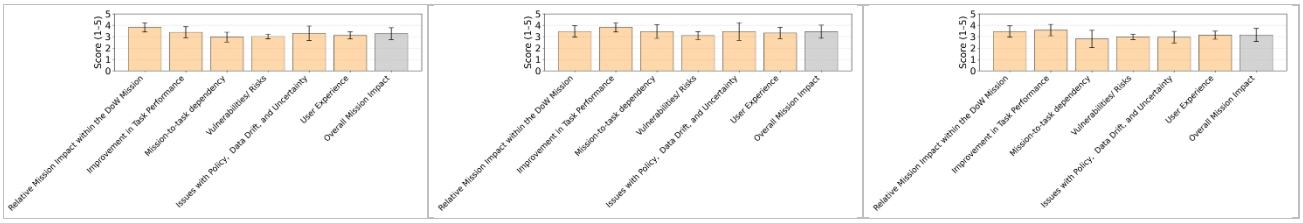


Figure 2. Mean and Standard Deviation Plots Across Use Cases

Comparison of Scores

Overall mission impact scores are more uniformly distributed, ranging from a mean of 3.12 for use case 1 to a mean of 3.43 for use case 5. The mission impact categories had a range of 0.31 on the 5-point scale. This compressed range for mission Impact scores indicates that the evaluation model scored the use cases with consistency despite implementation feasibility varying more. Use cases 3 and 4 achieved higher mission impact scores with averages of 3.29 and 3.26. Although use cases 3 and 4 scored higher in mission impact categories, they scored lower in implementation feasibility categories. When prioritizing use cases, a potential trade-off may be needed to select the most suitable use case where a higher mission impact score may carry greater implementation difficulties.

Statistical Analysis of LLM Evaluation

A three-factor factorial ANOVA was conducted for each of the 13 evaluation categories to assess the effects of use case, temperature, and evaluation stance on LLM-generated scores. Effect size was quantified using partial eta-squared (η^2_p), which measures the proportion of variance attributable to a given factor after accounting for other factors in the model. Partial eta-squared is defined in (1).

$$\eta_p^2 = \frac{SS_{\text{factor}}}{SS_{\text{factor}} + SS_{\text{error}}} \tag{1}$$

where SS_{factor} is the sum of squares associated with the factor of interest and SS_{error} is the residual sum of squares[18]. Thus, η^2_p represents the proportion of explainable variance in scores uniquely attributable to that factor.

Table 3. Effect and Significance of Use Case on Statistical Data

Capability	Use Case η^2_p	Use Case p-value	Effect Size	Significant p-value
Data Availability	0.32	0	Large	TRUE
AI Model Availability	0.40	0	Large	TRUE
Team AI Technical Expertise	0.55	0	Large	TRUE
Technology Maturity	0.51	0	Large	TRUE
System Complexity	0.03	0	Small	TRUE
User Training Requirements	0.40	0	Large	TRUE
Acquisition, T&E, and Approval	0.01	0.02	Negligible	TRUE
Relative Mission Impact within the DOW Mission	0.45	0	Large	TRUE
Improvement in Task Performance	0.24	0	Large	TRUE
Mission-to-task Dependency	0.20	0	Large	TRUE
Vulnerabilities/Risks	0.04	0	Small	TRUE
Issues with Policy, Data Drift, and Uncertainty	0.06	0	Medium	TRUE
User Experience	0.12	0	Medium	TRUE

For all categories, the test returned significant p-value, $p < 0.05$, for all use cases. This result indicates that the scores returned for separate use cases, although they may be very similar and have a lot of overlap, are statistically significant from one another. Effect size ranged across negligible effect $\eta^2_p < 0.01$, small effect $0.01 < \eta^2_p < 0.06$, medium effect $0.06 < \eta^2_p < 0.14$, and large effect, $0.14 < \eta^2_p$. This result suggests that the LLM can be used to give priority to use cases based on these categories as any differences, however small, are statistically significant.

Table 4. Effect and Significance of Temperature on Statistical Data

Capability	Temperature η^2_p	Temperature p-value	Effect Size	Significant p-value
Data Availability	0.03	0	Small	TRUE
AI Model Availability	0.01	0.03	Negligible	TRUE
Team AI Technical Expertise	0.01	0	Negligible	TRUE
Technology Maturity	0	0.44	Negligible	FALSE
System Complexity	0.01	0.01	Negligible	TRUE
User Training Requirements	0.01	0.01	Negligible	TRUE
Acquisition, T&E, and Approval	0	0.55	Negligible	FALSE
Relative Mission Impact within the DOW Mission	0	0.71	Negligible	FALSE
Improvement in Task Performance	0.01	0	Small	TRUE
Mission-to-task Dependency	0	0.76	Negligible	FALSE
Vulnerabilities/Risks	0.01	0	Negligible	TRUE
Issues with Policy, Data Drift, and Uncertainty	0.01	0.01	Negligible	TRUE
User Experience	0.02	0	Small	TRUE

Only some of the categories produced statistically significant data based on temperature. Even when the results were statistically significant, effect size for temperature only ranged negligible to small. This result suggests that temperature setting when running the LLM is largely unimportant, and settings to minimize errors are more important and are unlikely to affect the score outputs.

Table 5. Effect and Significance of Stance on Statistical Data

Capability	Stance η^2_p	Stance p-value	Effect Size	Significant p-value
Data Availability	0	0.37	Negligible	FALSE
AI Model Availability	0	0.45	Negligible	FALSE
Team AI Technical Expertise	0	0.48	Negligible	FALSE
Technology Maturity	0	0.99	Negligible	FALSE
System Complexity	0	0.78	Negligible	FALSE
User Training Requirements	0	0.31	Negligible	FALSE
Acquisition, T&E, and Approval	0.01	0.01	Negligible	TRUE
Relative Mission Impact within the DOW Mission	0	0.70	Negligible	FALSE
Improvement in Task Performance	0	0.55	Negligible	FALSE
Mission-to-task Dependency	0	1	Negligible	FALSE
Vulnerabilities/Risks	0	0.68	Negligible	FALSE
Issues with Policy, Data Drift, and Uncertainty	0.01	0.01	Negligible	TRUE
User Experience	0	0.92	Negligible	FALSE

Most categories did not produce statistically significant data based on stance. Stance also represented negligible effect on the variance in the data. This result suggests that stance setting when running the LLM is unimpactful, and the additional prompting related to setting the stance had no real effect on resulting scores produced.

Across all categories, use case was the dominant factor, exhibiting statistically significant effects ($p < 0.05$) in every case and medium-to-large effect sizes in most categories (η^2_p ranging from 0.031 to 0.550). This indicates that evaluation scores are primarily driven by substantive differences between use cases. Given this significance of data, the LLM can score use cases uniquely, and resulting scores can be used to make prioritization decisions as desired. However, these scores still need to be compared to human scoring to determine their accuracy.

RAG-LLM Scoring Comparison to Human Evaluation

The use cases scored by the RAG-LLM were also scored by a limited number of human evaluators to determine accuracy of the LLM's scores. Tables 6-11 show the comparison of RAG-LLM generated scores to human generated scores and compare the difference in average for each category. In only very few instances the RAG-LLM scores were close to the average score given by human evaluation, (within mean value of 0.20). This means the RAG-LLM did not regularly assign scores equal to human evaluators. However, this does not necessarily imply that the RAG-LLM is incorrect, as it may be scoring on a different scale than the human evaluators.

To examine the efficacy of the LLM's scoring, a ranking of the 6 use cases was created using the sum scores for the average of all LLM trials, and of the average of human trials. A rank is determined for cumulative overall implementation feasibility, mission impact, and combination of both. This rank, shown in tables 6 and 7, gives the prioritization that the LLM and the human evaluators would give to that use case. Although the LLM is seen to vary in scores dramatically from the human evaluators, it still selects the same use case as the human evaluators as best use case for investment.

Table 6. Rank Analysis for LLM Generated Scores

Category	Use case 1 RAG-LLM	Use case 2 RAG-LLM	Use case 3 RAG-LLM	Use case 4 RAG-LLM	Use case 5 RAG-LLM	Use case 6 RAG-LLM
Overall Implementation Feasibility	21.07	21.18	19.25	18.95	21.67	19.54
Overall Mission Impact	15.66	16.21	15.93	15.70	17.11	15.47
Overall Sum	36.73	37.39	35.18	34.65	38.78	35
Implementation Feasibility Rank	3	2	5	6	1	4
Mission Impact Rank	5	2	3	4	1	6
Overall RAG-LLM Rank	**3**	**2**	**4**	**6**	**1**	**5**

Table 6 displays the ranking that the LLM scores give to the six use cases. It ranks use case 5 as most worth investment, with use case 2 second.

Table 7. Rank Analysis for Human Generated Scores

Category	Use case 1 Human	Use case 2 Human	Use case 3 Human	Use case 4 Human	Use case 5 Human	Use case 6 Human
Overall Implementation Feasibility	23.25	23	23	22	26	21.50
Overall Mission Impact	16	18	16	15	19.50	16
Overall Sum	39.25	41	39	37	45.50	37.50
Implementation Feasibility Rank	2	3	3	5	1	6
Mission Impact Rank	3	2	3	6	1	3

Overall Human Rank	**3**	**2**	**4**	**6**	**1**	**5**

Table 7 summarizes the ranking that the human evaluation scores give to the six use cases. It ranks use case 5 as most worth investment, with use case 2 second. Importantly, the rank that the RAG-LLM gave to the 6 use cases was the same as the rank given by human evaluators.

Kendall τ-b statistics were computed to further compare human score results with RAG-LLM results. This test aligned each combination of use case and category RAG-LLM median score to its human median score counterpart. These pairs are then compared pairwise to each other pair to form the result[19]. The statistics are defined in the equation (2)

$$\tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}} \tag{2}$$

The calculation for the Kendall τ-b analysis was based on 78 paired observations derived from 13 categories across 6 use cases. These 78 pairs represent 3003 unique pairwise comparisons for calculation. The resulting Kendall τ-b was 0.288, with a p-value of 0.003. This indicates a statistically significant result, with a weak positive association between the human scores and RAG-LLM scores. This statistic suggests that higher human scores generally corresponded to higher RAG-LLM scores across categories.

5. CONCLUSIONS AND FUTURE WORK

The RAG-LLM capability scores for overall implementation feasibility range from 2.71 for use case 4 to 3.14 for use case 5. Average scores across all implementation feasibility categories spread across 0.43 points on the 5-point scale. Meanwhile, overall mission impact scores range from 3.12 for use case 1 to 3.43 for use case 5. The mission impact categories had a mean evaluation score range of 0.31 on the 5-point scale. Despite this small difference in means, statistical analysis indicates that RAG-LLM evaluation results for use cases are statistically distinct. Analysis also showed that temperature setting for the LLM had little to no impact on the resulting scores. Finally, direction of stance in prompting showed no impact on resulting scores.

The time to evaluate the categories for a single use case can take an hour for a human evaluation. The RAG-LLM capability can evaluate several use cases per minute, offering greater efficiency. Although the human evaluation data was limited, a comparison of the RAG-LLM scores to human scores showed that the LLM infrequently scored a category in each use case the same as the human evaluator. However, the LLM prioritized the use cases based on overall feasibility and mission impact scores in the same order as human evaluators. The LLM scores also showed a weak positive association with human scores. Since the relative use case prioritization is the goal of the RAG-LLM capability, this positive correlation of prioritizations shows a positive result and shows the ability to use RAG-LLM to score and prioritize use cases for AI investment.

Future Work

With this proof of concept, additional work could focus on refining the LLM to produce scoring results more in line with human evaluators. One such way could be to modify the materials provided to the LLM through RAG. Modifying these data sources based on use case specific data, agency specific data, or best practices may help to refine results. Additionally, more extensive human evaluation of a larger use case data set would help to contribute to tightening of LLM results and to assess the efficiency achievable over a human scoring. Broader human evaluation could help to show significance to the accuracy of the LLM results beyond their statistical independence.

ACKNOWLEDGEMENT

This work is supported by the U.S. Department of War (DOW) Developmental Test, Evaluation, and Assessments (DTE&A). Approved for Public Release; Distribution Unlimited. Public Release Case Number 26-0609.

DATA RIGHTS DISCLOSURE NOTICE

REFERENCES

[1] X. Mou, "Artificial intelligence: Investment trends and selected industry uses," *International Finance Corporation*, vol. 8, no. 2, pp. 311–320, 2019.

[2] J. Cheng, K. Ghate, W. Hua, W. Y. Wang, H. Shen, and F. Fang, "Realm: A dataset of real-world llm use cases," presented at the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 8331–8341.

[3] S. Jeon, Y. S. Chang, S. J. Jo, T. Madukuand, and Y. E. Kim, "Speed of catch-up and convergence of the artificial intelligence divide: AI investment, robotic, start-ups, and patents," *Journal of Global Information Technology Management*, vol. 27, no. 1, pp. 63–85, 2024.

[4] Y. Odeh and N. Al-Saiyd, "Prioritizing use cases: A systematic literature review," *Computers*, vol. 12, no. 7, p. 136, 2023.

[5] C. Catal and D. Mishra, "Test case prioritization: a systematic mapping study," *Software Quality Journal*, vol. 21, no. 3, pp. 445–478, 2013.

[6] R. Ahmed, D. Musleh, M. Ahmed, and M. El-Attar, "Use case prioritization using fuzzy logic system," presented at the 2014 IEEE 5th International Conference on Software Engineering and Service Science, IEEE, 2014, pp. 149–152.

[7] J. Liu, J. Lin, and Y. Liu, "How much can rag help the reasoning of llm?," *arXiv preprint arXiv:2410.02338*, 2024.

[8] M. L. Munoz and M. F. Islam, "Deep learning approaches for multi-class classification of phishing text messages," *Journal of Cybersecurity and Privacy*, vol. 5, no. 4, p. 102, 2025.

[9] T. Esho, C. Hoyt, J. Marshall, and J. Gadewadikar, "Artificial Intelligence Enabled Systems Engineering Modeling With Retrieval Augmented Generation," *Systems Engineering*, p. e70032, 2025.

[10] M. Fatehkia, J. K. Lucas, and S. Chawla, "T-rag: lessons from the llm trenches," *arXiv preprint arXiv:2402.07483*, 2024.

[11] I. Radeva, I. Popchev, L. Doukovska, and M. Dimitrova, "Web application for retrieval-augmented generation: Implementation and testing," *Electronics*, vol. 13, no. 7, p. 1361, 2024.

[12] M. Klesel and H. F. Wittmann, "Retrieval-augmented generation (rag) m. klesel, hf wittmann," *Business & Information Systems Engineering*, vol. 67, no. 4, pp. 551–561, 2025.

[13] H. Elkiran and J. Rasheed, "EvaRAG: Evaluating Advanced RAG Techniques With Indexing and Distance Metrics," *IEEE Access*, vol. 13, pp. 215724–215747, 2025.

[14] F. Pezoa, J. L. Reutter, F. Suarez, M. Ugarte, and D. Vrgoč, "Foundations of JSON schema," presented at the Proceedings of the 25th international conference on World Wide Web, 2016, pp. 263–273.

[15] A. Glentis *et al.*, "Scalable parameter and memory efficient pretraining for llm: Recent algorithmic advances and benchmarking," *arXiv preprint arXiv:2505.22922*, 2025.

[16] Z. Ságodi, I. Kolláth, P. Hegedűs, and R. Ferenc, "A Program Synthesis Dataset for LLM Temperature Analysis," *IEEE Access*, 2025.

[17] M. Renze, "The effect of sampling temperature on problem solving in large language models," presented at the Findings of the association for computational linguistics: EMNLP 2024, 2024, pp. 7346–7356.

[18] L. St and S. Wold, "Analysis of variance (ANOVA)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.

[19] P. K. Sen, "Estimates of the regression coefficient based on Kendall's tau," *Journal of the American statistical association*, vol. 63, no. 324, pp. 1379–1389, 1968.