Proceedings of the Annual General Donald R. Keith Memorial Conference
West Point, New York, USA
April 23, 2026
A Regional Conference of the Society for Industrial and Systems Engineering

Evaluating Retrieval-Augmented Generation for Academic Advising at the United States Military Academy

Andrew Miller and Donald Koban

Department of Systems Engineering, United States Military Academy, West Point, New York 10996

Corresponding author's Email: andrew.miller@westpoint.edu

Author Note: The views expressed herein are those of the authors and do not reflect the position of the United States Military Academy, the Department of the Army, or the Department of War.

Abstract: Modern colleges and universities have detailed curricula and policy requirements that make academic advising time-consuming and hard to scale. Large language models (LLMs) have the potential to automate routine advising tasks but often lack access to official institutional documents, limiting reliability in policy-driven advising contexts. Retrieval-augmented generation (RAG) has been proposed as a method to improve response grounding, but its performance in structured academic environments remains underexplored. This study evaluates the accuracy of a ChatGPT-based RAG system and Google's native AI search capability in answering common advising questions developed with faculty at the United States Military Academy. Responses were evaluated by a three-member grading panel using majority vote. ChatGPT augmented with institutional documents was correct 85% of the time, compared to 60% for Google's system. These findings suggest that institutions may benefit from maintaining custom RAG pipelines to improve advising accuracy in policy-constrained settings.

Keywords: Retrieval-Augmented Generation, Large Language Models, Academic Advising, Artificial Intelligence

1. Introduction

Academic advising plays a central role in helping students navigate course selection, degree requirements, and career planning . Traditionally delivered through in-person meetings, advising is time-intensive and dependent on individual counselor availability. As artificial intelligence tools become more widely adopted, institutions are exploring whether AI systems can support portions of the advising process. However, questions remain regarding the accuracy and reliability of AI-generated guidance in policy-driven academic environments.

Within the artificial intelligence space, retrieval-augmented generation (RAG) is particularly relevant to the requirements of academic advising because advising depends on institution-specific information -such as credit requirements, scheduling constraints, and internal policies- that is not always publicly available online. Retrieval-Augmented Generation (RAG) optimizes a large language model by having it reference specific data sources rather than relying solely on the general information a large language model (LLM) has learned during training (Su et al., 2025). A common limitation of LLMs is hallucination, in which the model generates false information when it lacks sufficient grounding. RAG systems mitigate hallucination by constraining responses to retrieved documents, thereby increasing the likelihood that answers are based on provided institutional data. While research has examined RAG models in academic advising contexts, one major gap remains: no RAG-based academic advising research has been conducted with respect to the United States Military Academy (USMA), at West Point.

USMA is a distinct advising environment because cadets must meet integrated requirements across academics, military development, physical fitness, and character, supported by departmental academic counselors who schedule academics alongside military and physical obligations. Most automated advising tools focus narrowly on academics and do not account for these multidomain constraints, making USMA a strong case for explainable, traceable systems such as retrieval augmented generation (Phillips et al., 2021). While advising chatbots are already used across universities to reduce bottlenecks and confusion (Amarnath & Nagarajan, 2024; Jayavardhana et al., 2025; Lugones et al., 2026; Tamascelli et al., 2025), evaluation remains central. RAG evaluation has been framed through context relevance, answer faithfulness, and answer relevance as binary classifiers (Saad-Falcon et al., 2023), alongside related metric sets such as Context Precision, Context Recall, Faithfulness, and Answer Relevancy, including baseline versus modified temperature comparisons (Amarnath & Nagarajan, 2024). Performance is sensitive to training and retrieval design choices, in that more training passages generally increase the accuracy of generative models (Izacard & Grave, 2021), and can be quantified using numerical scoring such as correctness and relevancy scores aligning outputs with relevant information (Mortaheb et al., 2025). These methods can be evaluated through pairing human ratings to assess agreement. Prior implementations illustrate feasibility: UWF's agentic ARGObot integrated

RAG with external tools (Serper, Gmail) and was evaluated on Answer Correctness, Context Precision, Faithfulness, Context Entity Recall, and Answer Relevancy (Tamascelli et al., 2025), while MyAdvisor used advisor interviews, observed sessions, scenario validation, and a Dialogflow based architecture, then evaluated learnability and usefulness with 17 students and 4 expert advisors, noting that about five participants often suffice to surface usability issues (Kuhail et al., 2023).

In this study, we systematically applied a RAG model within ChatGPT 5.2 to test a set of common prompts that reflected real predicaments DACs faced. We hypothesized that the model would demonstrate strong overall accuracy on advising questions but would struggle with complex scheduling questions due to limited access to highly specific scheduling constraints and cadet-context data. To test this hypothesis, we engaged with academic counselors at USMA to develop a set of salient questions. We then tested these questions using a RAG approach, supplying a ChatGPT model with USMA documents, course descriptions, and academic counseling guidelines. Through traceability, the model showed which documents it was pulling data from. Each response was evaluated by a three-grader panel, with majority vote determining the final score for each answer. Scores were then analyzed to assess model accuracy across the full question set and within key subcategories. To our knowledge, no prior studies had systematically applied retrieval-augmented generation to academic counseling at USMA or assessed its feasibility within the multidimensional requirements of a service academy. Our study found that although our model solved more fact-based questions easily, it struggled with complex procedural and scheduling problems, giving us direction towards future models to better accommodate counselors and cadets with complex advising concerns. Given this trend, our findings suggested that RAG models could benefit academic advising and might even support advisor-like functionality in the future, particularly if provided more complex data about the cadet experience alongside the institutional source documents already used for retrieval.

2. Methodology

2.1 Stakeholder Interview

To ensure the project scope aligned with real-world advising practices, we conducted a structured interview with the head Department Academic Counselor (DAC) for the Department of Systems Engineering. The DAC role is responsible for advising cadets on course selection, academic requirements, and departmental policies. During this interview, we documented the primary responsibilities associated with the DAC position and identified advising tasks that could potentially be supported by an artificial intelligence system. Particular attention was given to routine advising activities that require significant time investment from both cadets and counselors, such as scheduling meetings and answering common policy questions.

Insights from this interview informed the development of evaluation questions used to test the model. The stakeholder also agreed to assist in evaluating the model's responses to ensure that the assessment incorporated professional advising expertise in addition to the authors' evaluation based on official United States Military Academy (USMA) policy documents.

2.2 Model Development

After establishing advising requirements through the stakeholder interview, we developed a retrieval-augmented generation (RAG) model to answer advising-related questions. To maximize accessibility and ease of implementation, the model was implemented using OpenAI's ChatGPT platform. Although the platform does not provide a traditional RAG pipeline, a similar architecture can be created by uploading source documents that the model prioritizes when generating responses. OpenAI documentation notes that when knowledge retrieval is enabled and documents are uploaded, the system "dynamically retrieves relevant information" before generating a response, which corresponds to the retrieval-augmented generation (RAG) paradigm (OpenAI, 2025). Five official USMA advising and policy documents were uploaded to serve as the model's knowledge base:

1) USMA Redbook: Rulebook outlining all USMA courses, majors, and anything academic to include honors programs, pre-requisites, and special requirements.
2) USMA Greenbook: Rulebook outlining the USMA military program, where all military development courses are outlined.
3) USMA Whitebook: rulebook outlining the USMA Department of Physical Education (DPE) and the courses required for cadets, as well as physical graduation requirements.
4) USMA Advising Handbook: Advising handbook for USMA DAC's
5) Systems Department 8 Term Academic Plan (8TAP): General academic plans outlining the specific courses and engineering tracks for the two majors, systems engineering and systems decision sciences.

ISBN: 97819384962-5-7

These documents collectively capture the primary academic, military, and physical development requirements that guide cadet advising. By uploading these materials into the ChatGPT interface, the model was able to reference authoritative institutional documents when generating responses. Once these documents were incorporated, the model was ready for testing using representative advising questions.

2.3 Prompt Development and Model Testing

To evaluate the model's ability to respond to advising inquiries, we developed a set of 20 questions representing common scenarios encountered by academic counselors. Only 20 questions were created for this study because all responses were to be graded three times, and were projected to be lengthy to score, coming from a language model. The sample size of 20 questions was designed to cover all three pillars of cadet development -academic, military, and physical- while requiring the model to retrieve information from multiple documents. The prompts were organized into six categories reflecting different advising tasks: Simple / Fact-Based / Easily Verifiable (5 total), Simple but Conditional (3 total), Validations, Exceptions, Substitutes (4 total), Complex Procedural (3 total), Advising Judgement (2 total), Scheduling and Planning (3 total). Many prompts were intentionally designed to reference detailed policy information contained within the source documents in order to test the model's ability to retrieve specific rules.

To provide a baseline comparison, the same set of questions was also submitted to Google's native AI system integrated within the Google search engine. Unlike the RAG-based ChatGPT model, which relied on uploaded institutional documents as its primary knowledge source, Google's system generated responses using publicly available information retrieved from the internet. This comparison allowed us to evaluate whether a document-grounded advising model provides more reliable responses than a general-purpose AI search system.

All prompts and model responses were recorded in an Excel spreadsheet. Links to the original ChatGPT conversations were preserved to enable verification of responses and any follow-up interactions. Although follow-up prompts were occasionally used to examine whether the model corrected earlier responses, the primary focus of this study was the model's initial response, as this most closely reflects how cadets would interact with such a system in practice.

2.4 Score Model Prompts

After collecting responses from the model, we had to score our responses. In order to do this, we used the following standards to grade each response. Responses were graded on a scale of 1-3, from best to worst, respectively. The four classifications and descriptions are listed below:

1) Strongly Agree: Answer is fully correct by USMA standards.
2) Mostly Agree: Answer is mostly correct by USMA standards with few minor errors.
3) Incorrect: Answer has multiple minor errors or at least one major error that would misinform cadets and faculty about academic counseling and USMA guidelines.

Within these classifications, correctness will be determined based on the foundational documents. The three-point scale is useful instead of a two or four-point scale because the subjectivity across four responses would be difficult for graders to categorize, especially when LLMs often respond with long bodies of text. As for a two-point scale, the responses appeared too abstract to allow for a black-and-white categorization: "correct" or "incorrect". Instead, the three-point scale allows for the model to receive a high score if it perfectly answers the question, and if it is contextually right, but some parts of the response are partially incorrect, there is a classification for that. In our scale, minor errors include responses that don't explicitly answer the full part of the question or include aspects of a correct answer that have slight discrepancies with USMA rules. Major errors include responses that are fully incorrect with regard to feasibility and allowability. We scored the model, and to replicate the data collection, we had two USMA academic counselors score the model as well, with a total of 60 responses covering 20 questions. The prompt scores will then be displayed graphically and numerically to outline the sample breakdown of the model's ability to accurately score random questions.

3. Results

From a sample of 20 questions covering various USMA DAC concepts, we displayed the scores in a visual table using Excel. We then arranged the final scoring data to show proportions of the most common responses and visualized them through charts showing each scorer individually and both together. We scored 55% of the data as fully correct and another 30% as mostly correct. The remaining 15% were incorrect and too complex for the current model to solve. All incorrect problems fell

under the category of scheduling and planning. This was the only category that produced incorrect responses; all other question categories were either fully correct or mostly correct. The data from both our individual scoring and the combined scoring appear below in Table 1.

Table 1: The scoring breakdown for the GPT model next to Google's response score breakdown. The scoring breakdown for the GPT and Google is broken down by category on the 3-point scale, taken from our Excel data.

Question Category	**ChatGPT (n = 20)**				**Google (n = 20)**			
	Fully Correct	Mostly Correct	Incorrect	Mean Score	Fully Correct	Mostly Correct	Incorrect	Mean Score
Advising Judgement (n = 2)		2		2.00		2		2.00
Complex Procedural (n = 3)	3			1.00	2	1		1.33
Scheduling and Planning (n = 3)			3	3.00			3	3.00
Fact-Based / Easily Verifiable (n = 5)	2	3		1.60	1	1	3	2.40
Simple but Conditional (n = 3)	2	1		1.33	1	1	1	2.00
Validations, Exceptions, Substitutes (n = 4)	4			1.00	2	1	1	1.75
Total	11 (.55)	6 (.30)	3 (.15)	1.60	6 (.30)	6 (.30)	8 (.40)	2.10

The chart reflects that 85% of responses are relatively trustworthy based on our sample of questions. While acknowledging the benefits of this model, its ability to solve complex problems for scheduling and planning was the main reason it was unsuccessful. The implication of this is that scheduling and planning is a large piece of academic planning that would make this tool a real asset to cadets and faculty. By putting in the same Google questions, we found that the responses, although somewhat accurate, were much different in their composition, and many were significantly wrong. Google showed an overall 60% accuracy rate, where only 30% of all responses were fully accurate. Of the entire sample, 40% were incorrect. Evidently, Google underperformed compared to the model due to these differences, but these differences can be broken down further based on the questions' categories.

As shown in Table 2, Scheduling and planning received the worst average score as a category for both the GPT model and Google. The main difference between GPT and Google, however, was in the Simple / Fact Based category, where google failed to provide correct answers for all of those. In total, across all categories, on the 3-point scale, the average score of the GPT model was 1.65 on average, where the average Google score was 2.25. Therefore, Google tended to do worse than a 2 (mostly agreeable) for its responses. Because both systems were scored on the same set of prompts, we treated the data as paired observations and evaluated whether GPT achieved lower (better) rubric scores than Google using a one-sided paired, nonparametric Wilcoxon signed-rank test. This test is appropriate for our 1-3 ordinal scoring scale and does not assume normality of score differences over the sample size of 20. The analysis indicated that GPT scores were significantly lower than Google scores ($p = 0.0076$), supporting the conclusion that GPT outperformed Google on this evaluation set.

4. Discussion

The paired test shows the GPT based advising model is statistically more accurate than Google, reinforcing that performance is driven less by interface and more by data quality and control. Google's open web retrieval makes source selection largely uncontrollable unless users explicitly constrain it, which produced answers that were occasionally misinformed or incompatible with USMA standards and often required added context framing to avoid non-USMA assumptions. In contrast, the GPT model's retrieval augmented design grounds responses in selected USMA documents, improving traceability and institutional specificity, but it also exposes the central data dependency: its main failures occurred in scheduling and planning, where required constraints and procedures were not present in the provided sources. The implication is that accuracy gains come from curating and maintaining authoritative USMA data, especially structured scheduling guidance, rather than relying on general internet content.

GPT and Google differ most in source control, maintenance, and usability. Both can be accessed at little to no cost, but Google pulls from the open internet, so source quality is inconsistent unless users explicitly constrain it, which led to responses that were occasionally inaccurate or incompatible with USMA standards. A USMA tailored GPT would require periodic maintenance to keep its document set current as courses and policies change, but it offers stronger institutional specificity, higher accuracy in our results, and better usability because it can answer cadet and faculty questions without repeated context framing. Implementation cost would mainly vary with model sophistication and added capabilities such as APIs or external tool integrations, and both approaches provide some traceability that supports trust and accountability.

A digitized, always-available academic advising tool has clear potential benefits for cadet and faculty life at USMA, but effective adoption requires deliberate implementation measures. To address implementation responsibly, this study highlights the value of a sociotechnical approach: a framework for examining how technology interacts with human users, how it is used in practice, and the consequences of that use (Institute for Trustworthy AI in Law and Society (TRAILS), 2024). Sociotechnical approaches are increasingly applied across domains, especially for AI systems, because technical performance alone does not determine real-world success. For example, in healthcare delivery, sociotechnical framing is used to understand clinical decision support tools as complements to human judgment, improving decision-making in high-stress environments (Salwei & Carayon, 2022). While academic advising is less stress-inducing than clinical care, the underlying principle still applies: introducing an AI advising tool at USMA would likely shift advisor responsibilities from primarily informing and meeting toward verifying, validating, and ensuring alignment with institutional constraints. This transition would require training, updated workflows, and broad acceptance by faculty and advising personnel; without institutional alignment, the tool could introduce inconsistency or additional friction in the advising process. This emerging division of labor reflects the collaborative human-AI relationship central to sociotechnical models (Woodhams, 2026).

Although the GPT model frequently answered prompts correctly, several limitations should be addressed before considering implementation. First, the study used a limited question sample size. Because each response required hand grading, the sample had to remain manageable; this supports exploratory insight but limits generalizability. Second, we were constrained by the model tier and tooling available for this initial investigation. To keep the study feasible, we used the paid (plus) version of ChatGPT rather than higher-budget or enterprise configurations. As a result, the system could only incorporate a limited number of uploaded sources (five), which restricted coverage of niche but operationally important advising tasks. For example, the model lacked comprehensive support for building an 8TAP, which depends on precise formatting requirements and detailed, up-to-date constraints such as semester allotments, current capacity data, and other institution-specific rules. More advanced implementations could support richer data ingestion, structured constraint representations, and broader access to authoritative sources: capabilities that would likely improve accuracy and reduce failure modes in complex planning and scheduling scenarios.

5. Conclusion

This study indicates that a retrieval-augmented generation model implemented through ChatGPT can serve as a useful academic advising tool at USMA. In a baseline online configuration, the RAG approach outperformed Google in both response accuracy and ease of prompting. Although the model achieved 85% accuracy, its errors were concentrated in complex scheduling and planning tasks, areas not explicitly represented in the source materials and therefore likely addressable through the addition of authoritative scheduling procedures and constraint-based guidance.

Given its strong performance relative to Google, particularly without extensive user training or advanced prompt engineering, the RAG model demonstrates substantial potential for academic advising at USMA and for related applications. Because its responses are grounded in source documents, the system offers traceability while also supporting iterative clarification through dialogue. If implemented in its current form, the model would not replace departmental academic counselors, but it could shift routine advising interactions toward verification and exception handling. Such an approach could reduce repetitive back-and-forth communication and improve the efficiency of advising engagements for both cadets and faculty in a time-constrained environment.

With further development, including automated scheduling capabilities and broader system access, this approach could evolve into a more comprehensive academic advising solution. In that future state, human advisors would remain essential, but their role would increasingly center on oversight, judgment, and intervention in complex or exceptional cases.

6. Acknowledgements

The authors used ChatGPT to assist with editing this article by asking ChatGPT “Can you review this sentence or paragraph for clarity and conciseness?” ChatGPT offered recommendations that were incorporated into the final manuscript, after being provided with complete products prepared for review.

7. References

Amarnath, N. S., & Nagarajan, R. (2024). An Intelligent Retrieval Augmented Generation Chatbot for Contextually-Aware Conversations to Guide High School Students. *2024 4th International Conference on Sustainable Expert Systems (ICSES)*, 1393–1398. https://doi.org/10.1109/ICSES63445.2024.10762977

Institute for Trustworthy AI in Law and Society (TRAILS). (2024). *Response to NIST RFI on AI Executive Order 14110: The Importance of a Socio-technical Approach in AI Development*. Institute for Trustworthy AI in Law and Society (TRAILS).

Izacard, G., & Grave, E. (2021). *Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering* (arXiv:2007.01282). arXiv. https://doi.org/10.48550/arXiv.2007.01282

Jayavardhana, A., Hadinata, F. I., & Sanjaya, S. A. (2025). Optimizing Retrieval-Augmented Generation through Agentic RAG Ecosystem Based on Fine-Tuned BERT Cross Encoder and GPT-4 Model. *2025 IEEE International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, 1–7. https://doi.org/10.1109/AIMS66189.2025.11229773

Kuhail, M. A., Al Katheeri, H., Negreiros, J., Seffah, A., & Alfandi, O. (2023). Engaging Students With a Chatbot-Based Academic Advising System. *International Journal of Human–Computer Interaction*, *39*(10), 2115–2141. https://doi.org/10.1080/10447318.2022.2074645

Lugones, L. A. Q., Kverne, C., Bhimani, N. S., Oliveira, A. C., Polyzou, A., Lisetti, C., & Bhimani, J. (2026). *Aurora: Neuro-Symbolic AI Driven Advising Agent* (arXiv:2602.17999). arXiv. https://doi.org/10.48550/arXiv.2602.17999

Mortaheb, M., Khojastepour, M. A. A., Chakradhar, S. T., & Ulukus, S. (2025). *RAG-Check: Evaluating Multimodal Retrieval Augmented Generation Performance* (arXiv:2501.03995). arXiv. https://doi.org/10.48550/arXiv.2501.03995

Retrieval Augmented Generation (RAG) and Semantic Search for GPTs | OpenAI Help Center. (2025). OpenAI Help Center. https://help.openai.com/en/articles/8868588-retrieval-augmented-generation-rag-and-semantic-search-for-gpts?utm_source=chatgpt.com

Saad-Falcon, J., Khattab, O., Potts, C., & Zaharia, M. (2023). *ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2311.09476

Salwei, M. E., & Carayon, P. (2022). A Sociotechnical Systems Framework for the Application of Artificial Intelligence in Health Care Delivery. *Journal of Cognitive Engineering and Decision Making*, *16*(4), 194–206. https://doi.org/10.1177/15553434221097357

Su, W., Ai, Q., Zhan, J., Dong, Q., & Liu, Y. (2025). Dynamic and Parametric Retrieval-Augmented Generation. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 4118–4121. https://doi.org/10.1145/3726302.3731692

Tamascelli, M., Bunch, O., Fowler, B., Taeb, M., & Cohen, A. (2025). Academic Advising Chatbot Powered with AI Agent. *Proceedings of the 2025 ACM Southeast Conference*, 195–202. https://doi.org/10.1145/3696673.3723065

Woodhams, J. M. (2026). Emergent possibilities: A sociotechnical approach to generative AI in discourse analysis. *Research Methods in Applied Linguistics*, *5*(1), 100306. https://doi.org/10.1016/j.rmal.2026.100306