NPS-CM-24-221

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Evaluating SBIR Proposals: A Comparative Analysis using Artificial Intelligence and Statistical Programming in the DoD Acquisitions Process

June 2024

Maj Cullen G. Tores, USMC

Thesis Advisors: Dr. Maxim Massenkoff, Assistant Professor
Dr Robert F. Mortlock, Professor

Department of Defense Management

Naval Postgraduate School

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

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

The research presented in this report was supported by the Acquisition Research Program of the Department of Defense Management 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 Defense Management
Naval Postgraduate School

ABSTRACT

The Small Business Innovation Research (SBIR) program is a tool that the Department of Defense (DOD) uses to encourage industry development in technology that the market is otherwise not demanding. This helps to drive innovation and facilitate competition for government contracts. However, the source selection process within the SBIR program could be improved. It currently takes too long and is riddled with inconsistencies. Given this application and the rising interest in artificial intelligence (AI), it is worth exploring ways to augment the source selection process with AI. This study assesses the effectiveness of using large language models (LLMs) to automate classification of acquisition proposals as either competitive or noncompetitive. This study used R to extract text from the proposals, interact with OpenAI's models, and then iteratively loop through all of the proposals until completion. The intent was to establish a faster, more consistent, and objective evaluation system when compared to subjective human assessments. The final analysis indicated an emerging capability with vast potential, but one that is not reliable enough for immediate application into the SBIR program. This study emphasizes the importance of accuracy and reliability in DOD's initiatives and highlights the potential roles of AI in optimizing DOD acquisitions.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank the writing center for their unwavering support throughout my writing process. I want to specifically thank Chloe Woida for helping me refine my work from rough draft to final submission—no matter how many times I mistakenly thought the final draft was indeed "final." Her expertise as a writing coach was phenomenal, but I think the same could be said for all of the writing coaches. Chloe's compassion and commitment to getting me across the finish line was undeniable, and for that I'm very grateful.

The acknowledgments section would of course be incomplete if I failed to thank my advisors, Dr. Massenkoff and Dr. Mortlock. I'm both amazed and thankful that NPS was able to convince someone with Dr. Massenkoff's background and intellect to teach and advise within DDM. Dr. Massenkoff helped to take my original idea and develop it into something worthy of writing a thesis, even when it was uncomfortable. Dr. Mortlock helped me find my way into a dual degree program and has since been one of my academic advisors, and one of my professors for the better part of my time at NPS. Through it all, Dr. Mortlock has proven to be a professional that I truly aspire to be. His drive and integrity truly stand out among the staff, and I am sincerely grateful to have had him as an advisor through this process. Gentlemen, thank you.

Last but certainly not least, I want to thank my wife, Brittany. While I may have my shortcomings, you always find a way to help me deal with whatever it is I'm wrestling with. Your strength keeps me grounded when my mind wants to go astray, and for that I'm eternally grateful.

THIS PAGE INTENTIONALLY LEFT BLANK

NPS-CM-24-221

ACQUISITION RESEARCH PROGRAM
SPONSORED REPORT SERIES

Evaluating SBIR Proposals: A Comparative Analysis using Artificial Intelligence and Statistical Programming in the DoD Acquisitions Process

June 2024

Maj Cullen G. Tores, USMC

Thesis Advisors: Dr. Maxim Massenkoff, Assistant Professor
Dr Robert F. Mortlock, Professor

Department of Defense Management

Naval Postgraduate School

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

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AGI	Artificial General Intelligence
AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
BAAs	Broad Agency Announcements
CBA	Cost-Benefit Analysis
CSOs	Commercial Solutions Openings
DAU	Defense Acquisition University
DOD	Department of Defense
DSIP	Defense SBIR/STTR Innovation Portal
FN	False Negatives
FP	False Positives
FY	Fiscal Year
GPT	Generative Pre-trained Transformer
LLMs	Large Language Models
MARCORSYSCOM	Marine Corps Systems Command
MKSAP	Medical Knowledge Self-Assessment Program
OLS	Ordinary Least Squares
OPENAI	Open Artificial Intelligence
PCS	Permanent Change of Station
R&D	Research and Development
ROC	Receiver Operator Characteristic
SAT	Standardized Aptitude Test
SBIR	Small Business Innovation Research
STTR	Small Business Technology Transfer
SVM	Support Vector Machine
TN	True Negatives

TP	True Positives
TPOCs	Technical Points of Contact
U.S.	United States
UBE	Uniform Bar Examination

I. INTRODUCTION

The Small Business Innovation Research (SBIR) program is an important part of the Department of Defense's (DOD's) acquisition strategy. The program's objective is to foster innovative projects by providing funding to small businesses with research and development (R&D) projects that exhibit potential for commercialization (Held et al., 2006). The SBIR program injects innovation into the industrial base, fosters competition between government contractors, and is a tool to help shape technological advancements in areas that otherwise lack commercial incentive to advance. Despite the SBIR program's importance, the evaluation process—which is currently performed manually by human evaluators—is full of challenges and limitations. In the initial phase, SBIR invites small businesses to compete for awards focused on determining "the scientific and technical merit of proposed efforts" (Held et al., 2006). Promising projects progress to Phase II, which serves as the primary R&D phase, while Phase III entails further development, typically funded through private or other non-SBIR federal sources (Held et al., 2006). The practical problem with this system is the resource-intensive nature of manual evaluation, which requires significant time and effort from evaluators. Furthermore, inconsistencies may arise due to varying levels of subject matter expertise among evaluators, potentially leading to suboptimal evaluations. To address these issues, a streamlined and more consistent evaluation process is essential for enhancing the DOD's source selection efforts.

The current evaluation process presents practical problems due to its resource-intensive nature and potential for inconsistencies. Human evaluators are constrained by time and varying levels of expertise, and therefore may not apply the evaluation criteria consistently across different proposals. The pressure to review a large volume of proposals within strict timelines can often lead to rushed evaluations, with the potential for key aspects of a proposal's technical merit, professional qualifications, and potential for commercialization being overlooked. These issues could prevent the selection of promising projects that have the potential to enhance the military's capabilities. Despite

these challenges, the SBIR program's value is unquestionable. Audretsch et al. (2019) noted:

> Based on alternative evaluation methods applicable to survey data and case studies, we conclude that there is ample evidence that the DOD's SBIR Program is stimulating R&D as well as efforts to commercialize that would not otherwise have taken place. Further, the evidence shows the SBIR R&D does lead to commercialization, and the net social benefits associated with the program's sponsored research are substantial. (pp. 264–278)

This reinforces the SBIR program's worth and emphasizes the importance of a more efficient source selection process.

Additionally, a knowledge management problem in the source selection process exists. Technical points of contact (TPOCs) and evaluators, who possess varying levels of subject matter expertise relevant to the technical aspects covered by the SBIR topics, are the primary contributors to this gap. This knowledge deficit can exacerbate gaps in institutional knowledge, especially given the inherent volatility of human-based knowledge (Bollinger & Smith, 2001). The frequent rotation of TPOCs/evaluators before the completion of all SBIR phases further compounds this issue, potentially leading to an inadequate understanding of the technical aspects of the proposals. Within this context, artificial intelligence (AI), particularly through the use of large language models (LLMs) like ChatGPT, offer potential solutions. Guo et al. (2023) illustrated that ChatGPT provides objective answers and "generates safer, more balanced, neutral, and informative texts compared to humans" (p. 6). Furthermore, ChatGPT's capacity to "focus strictly on the given question" (p. 6) can help mitigate evaluation inconsistencies associated with the human evaluators' potential for divergent thinking. Modern technologies can be used to address this knowledge gap by integrating the benefits of human judgement with the objectivity and focus of AI platforms.

This study aims to bridge the gap in knowledge and offer a solution to the existing challenges in SBIR source selections by investigating the effectiveness of employing an LLM—specifically, the GPT-4 model from OpenAI, accessed via an application programming interface (API)—as a tool for assessment. OpenAI's GPT-4 model was

trained using supervised and reinforcement learning techniques on approximately 175 billion training parameters (George & George, 2023). By using one of the most advanced LLMs available for public use and evaluating its performance relative to human evaluators, this research provides insights into the accuracy, efficiency, and consistency of AI-based evaluations. This approach reduces the potential for human error and could lead to an improved overall evaluation process, all while potentially saving costs in the long term. In the way transformational innovations such as the internet or airplanes have reshaped society, the integration of advanced LLMs like OpenAI could signify a pivotal moment for the DOD's acquisitions process (George & George, 2023).

This study is guided by a set of primary and secondary research questions that focus on the proposal evaluation process and the potential benefits of integrating a generative text AI model.

The primary research question is

- How does the performance of automated evaluations using LLMs compare to that of human evaluators in the SBIR proposal evaluation process?

The secondary research questions are

- To what extent can automated evaluations using LLMs classify competitive and noncompetitive proposals?
- What are the potential broader implications of using automated evaluations using LLMs for optimizing source selection for contracts within the DOD acquisition effects beyond the SBIR program?
- What are the challenges and limitations associated with using an LLM in the SBIR proposal evaluation process, and how can they be addressed?

To answer the research questions, I will first explore prompt engineering literature and select three prompts that appear suitable to the task of proposal evaluation. I will then interact with OpenAI's GPT-4 model using R, a statistical programming tool, and an OpenAI API to automate proposal evaluation. After developing the required code, I will compare the distribution of human scores to the three prompt's scoring distributions. The

focus will then shift to regression analysis, where I will assess how the automated scores align with human assigned scores. Lastly, I will use the scores provided by automation to classify proposals as either competitive or noncompetitive. If the classification proves to be accurate and reliable, then evaluators would only have to review those proposals classified as "competitive" in detail. This has potential to reduce source selection time, allow human evaluators to focus their efforts more efficiently, and has potential to save money in terms of funding evaluation team members.

While this study intends to fully answer the research questions, there are certain limitations that should be clear. A review of the literature suggests that OpenAI's GPT-4 model is one of the most capable LLMs available for public use at the time of this study. However, OpenAI is fundamentally a commercial service, meaning that they are driven by profit, as are their competitors. Therefore, it is reasonable to assume that typical economic pressures and market competition will eventually result in increased LLM capability. More specific to this study, the output from this analysis will reflect the total number of proposals received, the three differing prompt engineering strategies, and the cumulative personal costs incurred through the study period. Put simply, more proposals or prompting strategies would lead to a more robust study but would increase personal costs.

This thesis addresses the details of SBIR source selection and potential automated solutions. Chapter I introduces the study, clearly outlining the research questions and identifying the potential role of LLMs in addressing the identified issues in SBIR proposal evaluation. After providing the necessary foundation, Chapter II offers a detailed overview of the SBIR program, followed by a review of the SBIR evaluation process. The chapter then pivots to a high-level overview of OpenAI LLMs and model comparisons, and their potential relevance to SBIR proposal evaluations. Chapter III explains the data and the methodology used in this analysis, followed by the analysis results in Chapter IV. This thesis concludes in Chapter V by offering a summary of findings and their implications, as well as a brief discussion about research limitations and recommendations for future research.

II. BACKGROUND AND LITERATURE REVIEW

This chapter begins by providing a background of the SBIR program and describes the current solicitation and source selection process. The focus then shifts into how modern technologies and LLMs could potentially optimize the evaluation process. Additional background information is provided for OpenAI's GPT performance and model comparisons. This OpenAI background concludes with a model-specific comparison of data usage and endpoint compatibility considerations. Finally, this chapter provides a review of the literature on AI-augmented classification and the emerging significance of prompt engineering in the context of LLM use.

A. SBIR BACKGROUND

The SBIR program serves as a mechanism for the federal government to influence technological development in areas that the commercial sector would otherwise not develop, mirroring the Berry Amendments' impact to the domestic textile industry. By focusing on small businesses, the SBIR program stimulates technological innovation and economic growth across a wide array of industries. The following sections offers an in-depth overview of the SBIR program at a functional level and provides a foundational understanding of the source selection process.

1. Overview of SBIR/STTR Programs

The SBIR program is important for stimulating technological innovation and economic growth within the country and operates in three primary phases. SBIR.gov's "About" page describes the three phases in SBIR:

> Phase I establishes the technical merit, feasibility, and commercial potential of proposed R&D efforts, with awards typically ranging from $50,000 to $250,000. Phase II progresses the R&D efforts initiated in Phase I and offers awards generally up to $750,000. Lastly, Phase III focuses on commercialization objectives resulting from the Phase I and II R&D activities, but it is not funded by the SBIR program. (2023)

Of note, specific funding thresholds are component-specific and sometimes include option periods. Option periods are generally only awarded if a vendor or

company is selected for Phase II and is intended to provide a funding buffer during the transition between Phase I and Phase II. Businesses are required to meet specific eligibility criteria for SBIR participation. This approach aims to direct the economic benefits towards small businesses to with the intent to foster innovation at the lowest levels and increase competition for government contracts.

2. SBIR Process and Evaluation

The DOD employs a robust SBIR proposal system to engage small businesses and research institutions to advance technology for defense applications. This proposal system is thoroughly documented by the DOD SBIR/STTR program (n.d.). The formal submission process is via the Defense SBIR/STTR Innovation Portal (DSIP), where DOD instructions guide applicants through both general and component-specific requirements. The DOD disseminates three standard Joint Broad Agency Announcements (BAAs) or Commercial Solutions Openings (CSOs) each fiscal year, with additional out-of-cycle or component-specific solicitations. BAAs and CSOs serve as the DOD's structured channels for acquiring R&D services and communicate capability needs to industry. The BAAs also specify the evaluation criteria for the vendors' awareness and guide the evaluation team in their assessments. The evaluation criteria and scoring methodology is listed in Table 1 and can be referenced at DOD SBIR/STTR Guide (n.d.). Criteria A addresses the technical aspects of a proposal and is the most heavily weighted among the three criteria, Criteria B emphasizes the qualifications and abilities of the team, while Criteria C assesses the commercial application for the technology.

Table 1. SBIR Evaluation Criteria. Adapted from DOD SBIR/STTR Guide (n.d.).

SBIR Evaluation Criteria		
Criteria	Definition	Points Scale
Criteria A	The soundness, technical merit, and innovation of the proposed approach and its incremental progress toward topic or subtopic solution.	0-40
Criteria B	The qualifications of the proposed principal/key investigators, supporting staff, and consultants. Qualifications include not only the ability to perform the research and development but also the ability to commercialize the results.	0-30
Criteria C	The potential for commercial (Government or private sector) application and the benefits expected to accrue from this commercialization.	0-30

After the vendor submission period ends, program offices receive completed proposals, and a TPOC is appointed by the program manager. Based on the Navy's "SBIR/STTR Topic Author and Topic Reviewer Guidebook" and my own personal experiences at Marine Corps Systems Command (MARCORSYSCOM), the TPOC would ideally be someone who was involved in requirement formulation. Specifically based on my experiences at MARCORSYSCOM, however, there are a variety of personnel considerations that actually influenced the TPOC assignment. These considerations included permanent change of station (PCS) timing, workload balance, and other personnel management considerations internal to the program team. Once assigned, the TPOC assembles an evaluation team based on the solicited requirements and the volume of submissions. This team conducts independent evaluations of proposals based on the predetermined evaluation criteria. Importantly, evaluators assess proposals on their own merits relative to the evaluation criteria rather than in comparison to each other, with only the final scores being compared for selection decisions—per the aforementioned guidebook. After evaluation, the TPOC compiles and normalizes the team assessments for a single submission to contracting.

B. OPENAI BACKGROUND

OpenAI has significantly sparked public interest in AI and has arguably reshaped the competitive landscape of AI usage since their introduction of ChatGPT. Larger companies have since developed their own LLMs in response to OpenAI's surging

popularity. Many experts in the field assert that we are experiencing the third "AI summer"—a term the AAAI Robert S. Engelmore Memorial Lecture describes as "a period of rapid scientific advances, broad commercialization, and exuberance—perhaps irrational exuberance—about our potential to unlock the secrets of general intelligence" (Kautz, H. 2022). The following sections provide the fundamental understanding of OpenAI necessary for this analysis, the different models available at the onset of this study, as well as important considerations for data usage and endpoint compatibility that will ultimately guide the study.

1. OpenAI and Multimodal LLM Overview

OpenAI was founded in 2015 and has elements of both a nonprofit and for-profit company. Their stated reason for this unique structure is to maintain their funding requirements while still facilitating the intellectual freedom of their employees. OpenAI's stated mission is to "develop artificial general intelligence (AGI) for the benefit of humanity," (OpenAI, n.d.). The November 2022 release of ChatGPT and GPT-3 brought OpenAI into the global spotlight, showcasing its capabilities in natural language processing.

The latest breakthrough from OpenAI is GPT-4, a significant evolution over its predecessors. GPT-4 is a multimodal language model that integrates both text and image inputs, which broadens its analytical and reasoning abilities. A detailed GPT-4 Technical Report available on OpenAI's website includes a table that presents these enhanced capabilities (see Figure 1).

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 (below 5th)	392 (below 5th)	260 (below 5th)
AP Art History	5 (86th - 100th)	5 (86th - 100th)	5 (86th - 100th)
AP Biology	5 (85th - 100th)	5 (85th - 100th)	4 (62nd - 85th)
AP Calculus BC	4 (43rd - 59th)	4 (43rd - 59th)	1 (0th - 7th)
AP Chemistry	4 (71st - 88th)	4 (71st - 88th)	2 (22nd - 46th)
AP English Language and Composition	2 (14th - 44th)	2 (14th - 44th)	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)	2 (8th - 22nd)	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)	5 (91st - 100th)	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)	5 (84th - 100th)	2 (33rd - 48th)
AP Microeconomics	5 (82nd - 100th)	4 (60th - 82nd)	4 (60th - 82nd)
AP Physics 2	4 (66th - 84th)	4 (66th - 84th)	3 (30th - 66th)
AP Psychology	5 (83rd - 100th)	5 (83rd - 100th)	5 (83rd - 100th)
AP Statistics	5 (85th - 100th)	5 (85th - 100th)	3 (40th - 63rd)
AP US Government	5 (88th - 100th)	5 (88th - 100th)	4 (77th - 88th)
AP US History	5 (89th - 100th)	4 (74th - 89th)	4 (74th - 89th)
AP World History	4 (65th - 87th)	4 (65th - 87th)	4 (65th - 87th)
AMC 10[3]	30 / 150 (6th - 12th)	36 / 150 (10th - 19th)	36 / 150 (10th - 19th)
AMC 12[3]	60 / 150 (45th - 66th)	48 / 150 (19th - 40th)	30 / 150 (4th - 8th)
Introductory Sommelier (theory knowledge)	92 %	92 %	80 %
Certified Sommelier (theory knowledge)	86 %	86 %	58 %
Advanced Sommelier (theory knowledge)	77 %	77 %	46 %
Leetcode (easy)	31 / 41	31 / 41	12 / 41
Leetcode (medium)	21 / 80	21 / 80	8 / 80
Leetcode (hard)	3 / 45	3 / 45	0 / 45

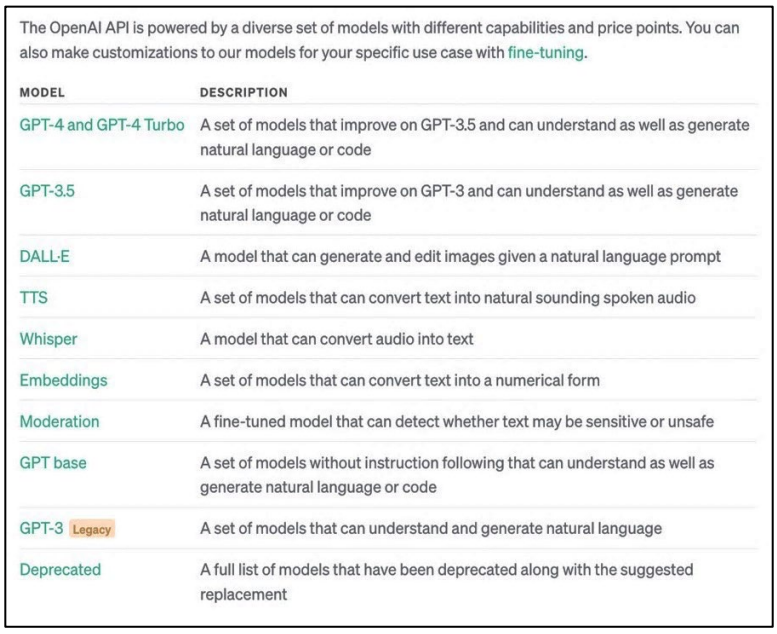
Figure 1. GPT-4 Technical Report Results. Source: OpenAI (2023).

The performance of GPT-4 and GPT-3.5 on the Uniform Bar Examination (UBE), the Standardized Aptitude Test (SAT) Evidence-Based Reading and Writing, and the Medical Knowledge Self-Assessment Program (MKSAP) aligns well with the diverse skill set required for evaluating SBIR proposals. Although the other tests might also offer valuable insights, the chosen exams provide a focused lens through which to gauge GPT-4's capabilities. GPT-4's impressive results, ranking in the top 10% on the UBE, top 7% on the SAT Reading and Writing, and top 25% on the MKSAP, highlight its advanced capabilities compared to GPT 3.5. These are key for comprehending and assessing complex SBIR proposals. GPT-3.5 also demonstrates competence, though to a lesser degree, scoring in the top 90% on the UBE, top 13% on the SAT Reading and Writing, and top 47% on the MKSAP. Given GPT-4's superior performance, it is likely to excel in

evaluating SBIR proposals. This hypothesis forms the basis of this study, which aims to explore GPT-4's capability in this context.

2. OpenAI API and Model Comparison

This analysis uses an OpenAI API key that is user-specific and enables the user to access specified models using a preferred software of choice. After obtaining an OpenAI API key, a user can then interact with the models listed in Figure 2 from OpenAI's publicly available API documentation.

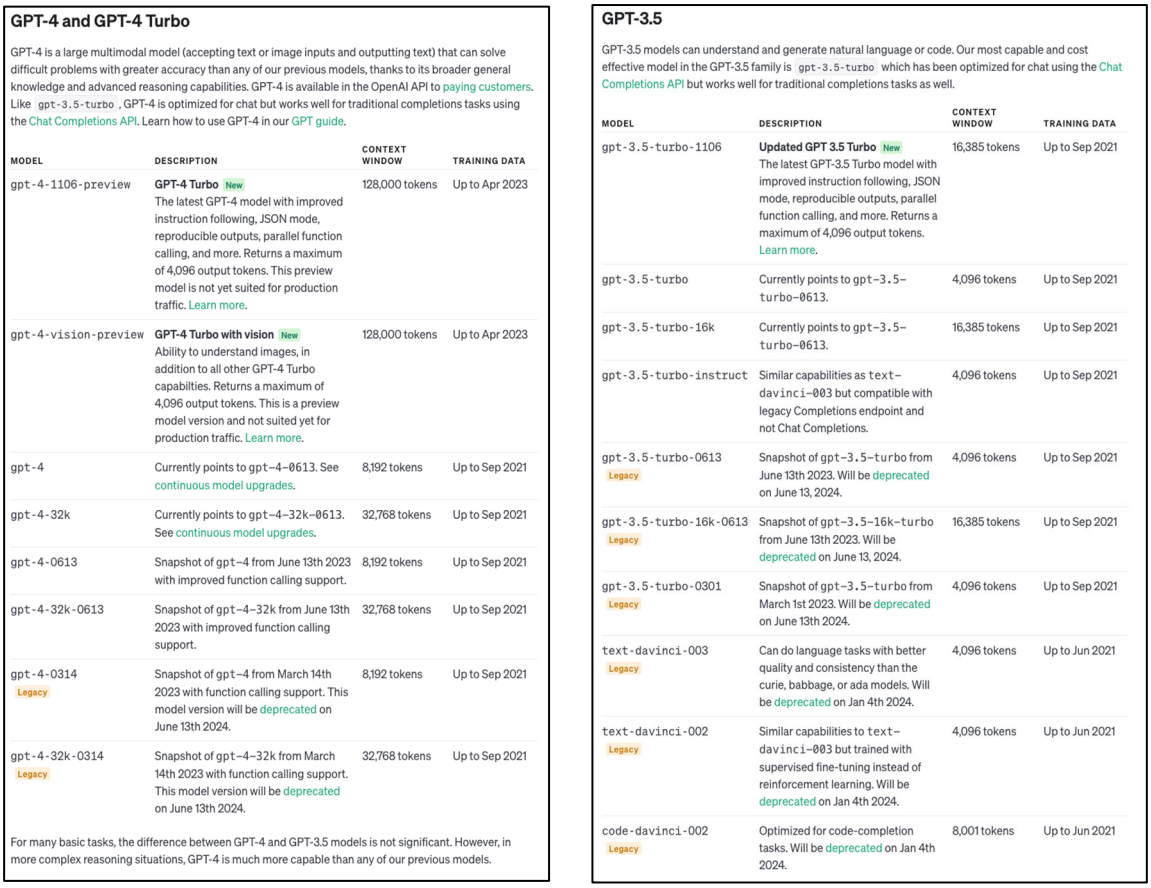


The OpenAI API is powered by a diverse set of models with different capabilities and price points. You can also make customizations to our models for your specific use case with fine-tuning.

MODEL	DESCRIPTION
GPT-4 and GPT-4 Turbo	A set of models that improve on GPT-3.5 and can understand as well as generate natural language or code
GPT-3.5	A set of models that improve on GPT-3 and can understand as well as generate natural language or code
DALL·E	A model that can generate and edit images given a natural language prompt
TTS	A set of models that can convert text into natural sounding spoken audio
Whisper	A model that can convert audio into text
Embeddings	A set of models that can convert text into a numerical form
Moderation	A fine-tuned model that can detect whether text may be sensitive or unsafe
GPT base	A set of models without instruction following that can understand as well as generate natural language or code
GPT-3 Legacy	A set of models that can understand and generate natural language
Deprecated	A full list of models that have been deprecated along with the suggested replacement

Figure 2. Model Descriptions. Source: OpenAI (n.d.).

Also listed on OpenAI's API documentation page are the GPT-4 model variants (as well as the legacy 3.5 and task-specific models such as DALL-E, which is used for text input and image output). I chose to use the GPT-4 model variant due to the performance comparisons mentioned above, the added benefits of a more recently updated training data (April 2023 versus September 2021), and the context window capacity of the GPT-4 models in comparison to the GPT 3.5 models. The side-by-side comparisons are shown in Figure 3.



GPT-4 and GPT-4 Turbo

GPT-4 is a large multimodal model (accepting text or image inputs and outputting text) that can solve difficult problems with greater accuracy than any of our previous models, thanks to its broader general knowledge and advanced reasoning capabilities. GPT-4 is available in the OpenAI API to paying customers. Like `gpt-3.5-turbo`, GPT-4 is optimized for chat but works well for traditional completions tasks using the Chat Completions API. Learn how to use GPT-4 in our GPT guide.

MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
`gpt-4-1106-preview`	**GPT-4 Turbo** New The latest GPT-4 model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. This preview model is not yet suited for production traffic. Learn more.	128,000 tokens	Up to Apr 2023
`gpt-4-vision-preview`	**GPT-4 Turbo with vision** New Ability to understand images, in addition to all other GPT-4 Turbo capabilties. Returns a maximum of 4,096 output tokens. This is a preview model version and not suited yet for production traffic. Learn more.	128,000 tokens	Up to Apr 2023
`gpt-4`	Currently points to `gpt-4-0613`. See continuous model upgrades.	8,192 tokens	Up to Sep 2021
`gpt-4-32k`	Currently points to `gpt-4-32k-0613`. See continuous model upgrades.	32,768 tokens	Up to Sep 2021
`gpt-4-0613`	Snapshot of `gpt-4` from June 13th 2023 with improved function calling support.	8,192 tokens	Up to Sep 2021
`gpt-4-32k-0613`	Snapshot of `gpt-4-32k` from June 13th 2023 with improved function calling support.	32,768 tokens	Up to Sep 2021
`gpt-4-0314` Legacy	Snapshot of `gpt-4` from March 14th 2023 with function calling support. This model version will be deprecated on June 13th 2024.	8,192 tokens	Up to Sep 2021
`gpt-4-32k-0314` Legacy	Snapshot of `gpt-4-32k` from March 14th 2023 with function calling support. This model version will be deprecated on June 13th 2024.	32,768 tokens	Up to Sep 2021

For many basic tasks, the difference between GPT-4 and GPT-3.5 models is not significant. However, in more complex reasoning situations, GPT-4 is much more capable than any of our previous models.

GPT-3.5

GPT-3.5 models can understand and generate natural language or code. Our most capable and cost effective model in the GPT-3.5 family is `gpt-3.5-turbo` which has been optimized for chat using the Chat Completions API but works well for traditional completions tasks as well.

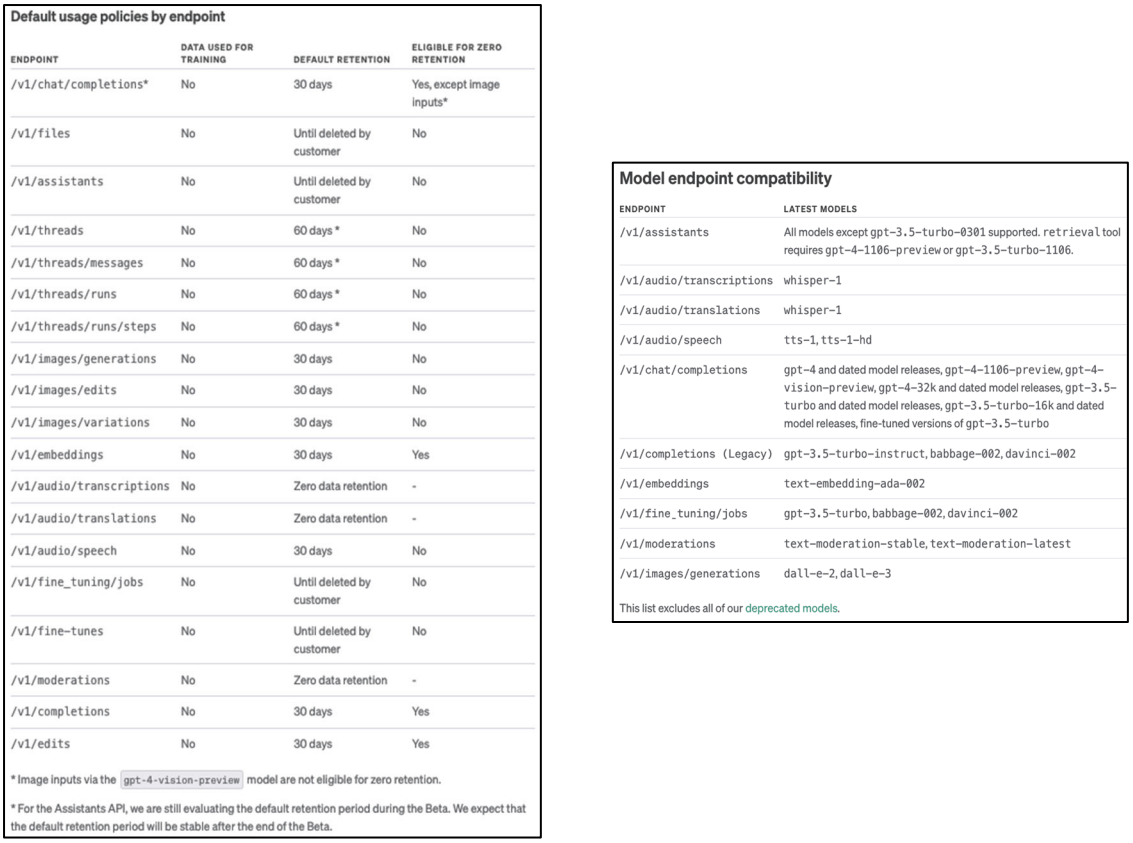
MODEL	DESCRIPTION	CONTEXT WINDOW	TRAINING DATA
`gpt-3.5-turbo-1106`	**Updated GPT 3.5 Turbo** New The latest GPT-3.5 Turbo model with improved instruction following, JSON mode, reproducible outputs, parallel function calling, and more. Returns a maximum of 4,096 output tokens. Learn more.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo`	Currently points to `gpt-3.5-turbo-0613`.	4,096 tokens	Up to Sep 2021
`gpt-3.5-turbo-16k`	Currently points to `gpt-3.5-turbo-0613`.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo-instruct`	Similar capabilities as `text-davinci-003` but compatible with legacy Completions endpoint and not Chat Completions.	4,096 tokens	Up to Sep 2021
`gpt-3.5-turbo-0613` Legacy	Snapshot of `gpt-3.5-turbo` from June 13th 2023. Will be deprecated on June 13, 2024.	4,096 tokens	Up to Sep 2021
`gpt-3.5-turbo-16k-0613` Legacy	Snapshot of `gpt-3.5-16k-turbo` from June 13th 2023. Will be deprecated on June 13, 2024.	16,385 tokens	Up to Sep 2021
`gpt-3.5-turbo-0301` Legacy	Snapshot of `gpt-3.5-turbo` from March 1st 2023. Will be deprecated on June 13th 2024.	4,096 tokens	Up to Sep 2021
`text-davinci-003` Legacy	Can do language tasks with better quality and consistency than the curie, babbage, or ada models. Will be deprecated on Jan 4th 2024.	4,096 tokens	Up to Jun 2021
`text-davinci-002` Legacy	Similar capabilities to `text-davinci-003` but trained with supervised fine-tuning instead of reinforcement learning. Will be deprecated on Jan 4th 2024.	4,096 tokens	Up to Jun 2021
`code-davinci-002` Legacy	Optimized for code-completion tasks. Will be deprecated on Jan 4th 2024.	8,001 tokens	Up to Jun 2021

Figure 3. Model Tokens and Training Update. Source: OpenAI (n.d.).

An important consideration in model selection worth more discussion is the content window and the concept of tokens. All the available models break down and process text into tokens. While OpenAI's documentation page does not provide a formulaic breakdown showing how tokens are calculated, it does say, "As a rough rule of thumb, 1 token is approximately 4 characters or 0.75 words for English text" (OpenAI, n.d.). This directly impacts the capacity of the model to receive a proposal as input and then provide a text-based output because the sum of both input and output cannot exceed the token limitation of the specified model. The model with the largest token context window—and therefore the most capable model for proposal inputs—was the "gpt-4-1106-preview" model.

3. Data Usage and Endpoint Compatibility

SBIR programs typically encompass innovative R&D projects that result in the creation of valuable intellectual property and proprietary information. Ensuring the secure management and protection of such sensitive data is critical, as it preserves the investments of both the government and the small businesses involved. Notably, there is a distinction in the handling of data by OpenAI services: Data submitted to OpenAI's non-API consumer services, including ChatGPT, is retained and may be utilized to further train OpenAI's machine learning models. In contrast, data interfaced with OpenAI models through the API is not incorporated into training data sets. This differentiation is significant for this analysis to proceed, specifying that putting SBIR-related data into non-API services might not conform to the security protocols necessary for safeguarding proprietary information. The screenshots provided from OpenAI's API documentation confirm that data from API services are not used for model training, illustrate the default data retention periods (which serve solely for abuse prevention), verify the option for zero retention eligibility, and demonstrate compatibility with the "gpt-4-1106-preview" and the "/v1/chat/completions" endpoint (see Figure 4).



Default usage policies by endpoint

ENDPOINT	DATA USED FOR TRAINING	DEFAULT RETENTION	ELIGIBLE FOR ZERO RETENTION
/v1/chat/completions*	No	30 days	Yes, except image inputs*
/v1/files	No	Until deleted by customer	No
/v1/assistants	No	Until deleted by customer	No
/v1/threads	No	60 days *	No
/v1/threads/messages	No	60 days *	No
/v1/threads/runs	No	60 days *	No
/v1/threads/runs/steps	No	60 days *	No
/v1/images/generations	No	30 days	No
/v1/images/edits	No	30 days	No
/v1/images/variations	No	30 days	No
/v1/embeddings	No	30 days	Yes
/v1/audio/transcriptions	No	Zero data retention	-
/v1/audio/translations	No	Zero data retention	-
/v1/audio/speech	No	30 days	No
/v1/fine_tuning/jobs	No	Until deleted by customer	No
/v1/fine-tunes	No	Until deleted by customer	No
/v1/moderations	No	Zero data retention	-
/v1/completions	No	30 days	Yes
/v1/edits	No	30 days	Yes

* Image inputs via the `gpt-4-vision-preview` model are not eligible for zero retention.

* For the Assistants API, we are still evaluating the default retention period during the Beta. We expect that the default retention period will be stable after the end of the Beta.

Model endpoint compatibility

ENDPOINT	LATEST MODELS
/v1/assistants	All models except `gpt-3.5-turbo-0301` supported. `retrieval` tool requires `gpt-4-1106-preview` or `gpt-3.5-turbo-1106`.
/v1/audio/transcriptions	whisper-1
/v1/audio/translations	whisper-1
/v1/audio/speech	tts-1, tts-1-hd
/v1/chat/completions	gpt-4 and dated model releases, gpt-4-1106-preview, gpt-4-vision-preview, gpt-4-32k and dated model releases, gpt-3.5-turbo and dated model releases, gpt-3.5-turbo-16k and dated model releases, fine-tuned versions of gpt-3.5-turbo
/v1/completions (Legacy)	gpt-3.5-turbo-instruct, babbage-002, davinci-002
/v1/embeddings	text-embedding-ada-002
/v1/fine_tuning/jobs	gpt-3.5-turbo, babbage-002, davinci-002
/v1/moderations	text-moderation-stable, text-moderation-latest
/v1/images/generations	dall-e-2, dall-e-3

This list excludes all of our deprecated models.

Figure 4. Data Used for Training. Source: OpenAI (n.d.).

C. LITERATURE REVIEW ON AI-AUGMENTED CLASSIFICATION AND PROMPT ENGINEERING

A review of the literature demonstrates that real world applications of AI-augmented technologies is proving beneficial in classification tasks, particularly in medical fields. These sections detail AI classification studies related to the diagnosis of COVID-19 and colorectal cancer screenings—both of which saw benefits in increased accuracy and speed. The specific AI tool of interest for this study is LLMs. Current literature found that specific word choice in prompts has potentially significant impacts to the quality of the response. The following sections discuss the emerging importance of prompt engineering and an existing catalog of prompt strategies that facilitate this study.

1. AI-Augmented COVID-19 Detection

Ghayvat et al.'s 2022 study found that chest computerized tomography (CT) scans could be an important complementary tool in diagnosing COVID-19 through AI-supported image classification. The standard diagnostic method had demonstrated limitations in availability and turnaround time relative to COVID-19's spread. The primary approach that the study introduced was the "Radiologist in the Loop" (RIL) model, which integrated human expertise with AI capabilities. The study ultimately found that the RIL model had similar accuracy measures as unassisted radiologist (since the RIL ensured a human radiologist made the final classification determination) but dramatically decreased the diagnosis time. The traditional manual diagnosis process was listed at 225.5 minutes, while the RIL model decreased that time to approximately 7 minutes after a just few model updates. While the accuracy metrics were largely similar, this dramatic decrease in diagnosis time indicates a distinct increase in efficiency, which was particularly important considering the context of minimizing the spread COVID-19. This study provides a unique perspective on AI-augmented classification, shifting the focus from the conventional accuracy-driven outcome to an efficiency-driven outcome, which may be applicable in certain DOD acquisition contexts.

2. AI-Augmented Colorectal Polyp Classification

Nasir-Moin et al.'s 2020 study revealed that AI-augmented techniques outperformed traditional microscopic assessments in colorectal cancer screenings, increasing the accuracy from 73.9% to 80.8%, and even achieved 87% accuracy when using AI alone. The study specifically assessed colorectal cancer screenings, the accuracy of the screenings, and the impacts to follow up exam scheduling. Nasir-Moin points out issues both with scheduling too many follow-up appointments as well as not enough, ranging from patient inconvenience, inflated health care costs, and the potential for negative cancer outcomes. The study's intent for augmenting this process with AI was to improve the accuracy of these screenings. When assessing the traditional screening methods, Nasir-Moin et al. note that the variation in human pathologists will naturally be cause for some of the error. Additionally, the study notes a growing shortage of

pathologists and suggests the possibility of delayed and inaccurate screenings resulting from being overworked. The study concludes by suggesting that the widespread implementation of such an automated classifier would save time, money, and produce better health outcomes. The evidence from Nasir-Moin et al.'s work highlights the potential of AI and supports further exploration of AI tools in a variety of use cases.

3. Prompt Engineering and Optimization

The study *Large Language Models as Optimizers* by Chengrun Yang et al. (2023) emphasizes the critical role of prompt engineering in getting the highest quality responses from large language models. A noteworthy case from the paper illustrates the nuance of prompt design with an experiment involving an LLM created by Google on a grade school–level math test. They found that the prompt "Let's think step by step" achieved a relatively high accuracy of 71.8. A slightly different prompt, "Let's solve the problem together," resulted in a lower accuracy of 60.5. Interestingly, when these two prompts were semantically merged into "Let's work together to solve this problem step by step," the accuracy further decreased to 49.4. This counterintuitive outcome highlights how subtle variations in prompt composition can lead to dramatic changes in the performance of LLMs. The case illustrates the emerging importance of prompt engineering and highlights the sensitive and complex dynamics between the prompt's structure and the AI's resulting performance.

The significance of prompt engineering is further elaborated on in White et al.'s (2023) *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. This comprehensive guide provides a structured approach to prompt design, offering a catalog of 16 prompt patterns that seek to optimize the interaction with conversational LLMs. It reinforces that the choice of prompt significantly influences the model's performance and then attempts to provide a framework for structuring prompts to specified use-cases. The catalog presents various patterns with their specific purposes, structures, and examples. This enables users to develop their interaction strategies based off their specific use case. This framework was particularly useful in structuring prompts that facilitate automated SBIR proposal evaluations. While these prompt patterns offer a foundation for LLM

interactions in general, in practice there is SBIR-specific content that must be included in the prompt to ensure the automated process aligns with SBIR's requirements. Specifically, this includes integrating the mandatory evaluation criteria and the topic requirements that each proposal is evaluated against.

D. APPLICATION TO THIS STUDY

The preceding information provides the necessary frameworks of the SBIR program to understand how and where the SBIR source selection process can be improved. The chapter then discussed emerging capabilities and considerations with LLMs, and then demonstrated through the literature various instances of successful AI augmentation. Through this lens, this study uses a personal API key to interface with the "/v1/chat/completions" endpoint, which then communicates with the GPT-4 model for increased performance relative to GPT-3.5. The study recognizes the significance of prompt structure and assesses three different prompt strategies to compare performance. The first prompt is a customized, user created prompt; the second prompt uses a flipped interaction approach based off the literature; and the third prompt uses an adopted persona approach from the literature.

III. DATA AND METHODOLOGY

This chapter presents the core elements of the study. It begins with a discussion of the data sources and the compilation process, which used statistical programming tools to integrate with OpenAI via API. The chapter then explains the prompts chosen for the analysis and the reasoning behind their selection. Additionally, this chapter provides an overview of the dataset, highlighting the varying numbers of observations by prompt and the details of the variables involved. The chapter concludes by summarizing the analytical methods employed in the research.

A. DATA SOURCE

The main data for this analysis are final consensus scores that were assigned by TPOCs and their evaluation teams from MARCORSYSCOM. All the proposals analyzed in this study were selected by the MARCORSYSCOM SBIR program manager, and appropriate vetting through MARCORSYSCOM Legal Office was conducted by the program manager prior to distribution. The only limitation imposed on this study by the Legal Office was that specific company names and proposal numbers could not be published or otherwise made public. The imposition of this restriction did not affect the scope of this analysis. This study used a data set of 133 proposal evaluations from BAA cycles 17.1, 18.1, and 18.2.

B. DATA COMPILATION

The individual criteria and total scores for each proposal were initially manually compiled into a Microsoft Excel file. Then, automated scores from OpenAI's "gpt-4-1106-preview" model were required for final compilation into data sets ready for analysis. This was completed using a statistical programming software called "R" through a detailed script that interfaced with a personalized API key from OpenAI. This script first retrieved predefined topic descriptions and requirements, followed by text preprocessing functions to the proposal texts. The preprocessing functions standardized case, removed irrelevant characters, and removed common words that did not contribute to the evaluation. This also helped to minimize the token count to increase processing

speed and reduce personal cost. The script then generated scores, recorded the results into a text file, and extracted these into a separate data frame for analysis. This process was repeated with three distinct prompts to evaluate the impact of prompt engineering on the effectiveness of the automated scoring system.

C. PROMPT DESIGN

In this study, three distinct prompts were developed: one was self-composed, another was developed using the "flipped interaction" pattern for a reversed perspective, and the third employed the "persona" pattern, which instructed the model to embody the persona of a seasoned SBIR evaluator. The self-composed prompt encompasses the evaluation criteria verbatim from Table 1 and can be externally referenced at DOD SBIR/STTR Guide (n.d.). The other two prompts define the evaluation criteria as "evaluation_criteria" in R's global environment and then reference "evaluation_criteria" through code versus spelling the criteria out manually. Two prompt patterns were selected from the catalog of 16 due to their applicability to this use-case. The "flipped interaction" pattern elicits a role reversal that engages the LLM to generate its own questions deemed necessary to generate an accurate response. The "persona" pattern creates a role-play scenario, where the LLM takes on the role of a character with expertise specific to the role given. All three prompts are shown in Table 2.

Table 2. Prompt Composition

Analysis Prompts		
	Prompt Structure	Prompt Logic
Prompt 1	Critically evaluate the SBIR proposals with the understanding that only one or two will be selected for a Phase I award. This competitive context means that while each proposal may have strengths, they must be distinguished from each other with a realistic spread of scores. Not all proposals can score highly unless they stand out with exceptional qualities. Evaluation Criteria: a. The soundness, technical merit, and innovation of the proposed approach (0–40 points). Provide a critical assessment, ensuring that high scores are justified by exceptional factors that clearly differentiate the proposal from others. b. The qualifications of the proposed principal/key investigators (0–30 points). Scores should reflect the ability of the investigators to stand out in a competitive field, with specific strengths and weaknesses that could realistically impact their chances of selection. c. The potential for commercial application and benefits of commercialization (0–30 points). Evaluate with a critical eye on market potential and the challenges faced, understanding that only the most viable and well-positioned proposals will be selected." Topic Requirements: topic_requirements Proposal: proposal_text Provide a detailed score for each of the three criteria, including justifications for points awarded or deducted. Scores should reflect the competitive goal of selecting only the most promising proposals for a Phase I award. Always output the scores in the exact same format at the conclusion of each evaluation. ### Criteria A Score: XX / 40, ### Criteria B Score: XX / 30, ### Criteria C Score: XX / 30	Personally Drafted
Prompt 2	Begin by briefly summarizing the SBIR proposal: Proposal: proposal_text Next, using the flipped interaction method, critically evaluate the proposal by asking detailed questions based on the following criteria: evaluation_criteria For each criterion, formulate questions that explore the strengths, weaknesses, and unique aspects of the proposal. Consider the competitive context of the SBIR program and the necessity for proposals to stand out. Examples of questions include - For Criterion a, ask about specific innovative elements and how they surpass standard approaches in the field. - For Criterion b, inquire about the unique qualifications and past achievements of the key investigators. - For Criterion c, question the commercialization strategy and potential market challenges. After asking these questions, evaluate the proposal based on the answers. Assign a numerical score out of the total points available for each criterion. Use the following guidelines for scoring: - Score high (close to the maximum points) if the proposal demonstrates exceptional innovation, qualifications, or market potential. - Score medium (around the midpoint of the scale) if the proposal is solid but lacks outstanding qualities. - Score low (towards the lower end of the scale) if there are significant concerns or	Flipped Interaction

Analysis Prompts		
	Prompt Structure	Prompt Logic
	shortcomings. Scores should reflect the depth of innovation, team qualifications, and commercial potential as revealed through the evaluation. Avoid using placeholders or leaving scores as TBD. Conclude with a detailed evaluation and scoring for each criterion, justifying the scores with insights gained from the flipped interaction questions. Output the scores in this format: "### Criteria A Score: [Exact numerical score out of 40] – Justification: [Brief explanation]" "### Criteria B Score: [Exact numerical score out of 30] – Justification: [Brief explanation]" "### Criteria C Score: [Exact numerical score out of 30] – Justification: [Brief explanation]" Reminder: Each score must be a specific numerical value, directly reflecting the proposal's strengths and weaknesses in each criterion. Please avoid ambiguous or incomplete scoring.	
Prompt 3	As a seasoned and critical SBIR evaluator with extensive experience, approach the evaluation of the following SBIR proposal methodically. Your expertise lies in not only identifying the strengths and weaknesses of a proposal but also in discerning its potential for success in a competitive environment. Take your time to think through each aspect of the proposal, reflecting on your vast experience and using a chain of thought process to arrive at a well-reasoned evaluation. Begin with a brief summary: Proposal: proposal_tex Engage in a detailed chain of thought process as you critically evaluate the proposal against the following criteria: evaluation_criteria - Consider each aspect carefully, weighing its merits and potential pitfalls. - Reflect on similar proposals you've evaluated in the past and draw comparisons where relevant. - If unsure about a particular aspect, reason it out step-by-step, just as you would approach a complex calculation. - After each criterion, pause to ask yourself if there's anything you might have missed or any additional insight you can apply. Use the following guidelines for scoring: - Clear differentiation is key: score exceptional proposals in the high range (85–100), solid proposals in the midrange (65–84), and noncompetitive ones lower (<65). - Provide a specific numerical score for each criterion, with justification. Finally, compile your scores and justifications, ensuring they reflect a critical and fair assessment: "### Criteria A Score: XX / 40, ### Criteria B Score: XX / 30, ### Criteria C Score: XX / 30" Remember: Your goal is to discern between potentially successful and noncompetitive proposals, focusing on overall success potential in the SBIR program's competitive context.	Adopted Persona
*Indicates that unlike Prompt-1, this prompt has the evaluation criteria pre-defined in R's global environment as "evaluation_criteria"		

There were instances where the prompting failed. Each data set ultimately contained a different number of observations based on feedback from the OpenAI model. For example, when each proposal was evaluated against Prompt 2, 16 proposals returned an incomplete evaluation. Similarly, when using Prompt 1, there was one proposal that returned an incomplete evaluation. The third prompt was the only one to return all 133 evaluations. In addition to these prompt anomalies, there was also a degree of variation in scoring output, despite their being explicit instructions within each prompt. This affected the automated score extraction process, ultimately requiring a supplemental manual compilation.

D. DATA DESCRIPTION

Three different data sets were used for this analysis to compare the differing prompting strategies. The Prompt 1 data set contained 132 observations, the Prompt 2 data set contained 117 observations, and the Prompt 3 data set contained 133 observations. This variance in observation count is attributed to the prompt anomalies described in the Data Compilation section. Each data set consisted of 16 variables, which are described in detail in Table 3.

Table 3. Variable Name and Description

Variable Name	Variable Description
X	A number variable that simply counts the number of proposals in sequential order.
p_number	This represents the specific proposal number
t_number	This represents the topic number and facilitated grouping
h_a_score	Human evaluation score for evaluation criteria "A"
h_b_score	Human evaluation score for evaluation criteria "B"
h_c_score	Human evaluation score for evaluation criteria "C"
h_t_score	The total score assigned by human evaluation team
c_a_score	Computer evaluation score for evaluation criteria "A"
c_b_score	Computer evaluation score for evaluation criteria "B"
c_c_score	Computer evaluation score for evaluation criteria "C"
c_t_score	The total score assigned by OpenAI model
c_t_score_decile	This represents the computer evaluation scores broken into deciles. This facilitated visual representation of the relationship between computer total scores and human total scores
percentile_25_h	This variable is the calculated 25th percentile of the human evaluation total scores.
percentile_25_c	This variable is the calculated 25th percentile of the computer evaluation total scores.
h_noncompetitive	A binary variable where 1=proposal less than the 25th percentile of human total scores, 0=greater than the 25th percentile. This variable tells us whether or not the proposal is non-competitive (below the 25th percentile).
c_noncompetitive	A binary variable where 1=proposal less than the 25th percentile of computer total scores, 0=greater than the 25th percentile. This variable tells us whether or not the proposal is non-competitive (below the 25th percentile)

E. METHODOLOGY AND MODELS

The initial step in this analysis was to examine the distribution of both human and computer-generated scores across all three prompts. This step was important because it highlighted potential scoring biases and outliers that served to inform the study's interpretation. Afterwards, I studied the relationship between human and computer-generated scores.

Ordinary least squares (OLS) regression was used to compare human scores and computer-generated scores. The OLS models used in this portion of the study are represented by the following univariate equation:

$$Y = \beta 0 + \beta 1\, X + \varepsilon$$

In this formula, the human total score is represented by "Y," while the computer-generated score is represented by "X." The "β_1" term represents the coefficient or slope of the line, which is the expected change in "Y" given a one-unit change in "X." "β_0" represents the predicted human total score when the computer-generated score is equal to zero, and "ε" represents the error term.

After regression analysis, I assessed the potential for computer-generated evaluations to classify proposals as either competitive or noncompetitive. This reorientation sets up a binary classification task. The purpose of this classification is to save evaluators time and effort by reducing the number of proposals that they would have to review in detail. To do this, I first observed the distribution of computer scores relative to human classification of competitive and noncompetitive. Importantly, evaluators do not formally classify proposals in this manner in practice. I computed the classification thresholds based off the average value of the lowest quartile of human-generated scores. I used the average value of the lowest quartile of human scores because scores varied based on topic, who the evaluators were, and any number of things beyond the scope of this analysis. The distribution of computer scores relative to human classification informed the final area of study, where the prompts' classification accuracy was determined using confusion matrices. The study concludes with a receiver operator characteristic (ROC) curve plot that illustrates the overall performance of each prompt and represents all the varying degrees of true and false positive ratios.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. RESULTS

This chapter details the outcomes of the study. It begins by presenting the descriptive statistics of the data, followed by an explanation of the regression analysis that examine scoring alignment. Next, this chapter evaluates the ability to classify proposals as either competitive or noncompetitive, compares confusion matrices to assess the accuracy of predictions, and analyzes ROC curves to measure overall model performance. Finally, the chapter concludes with a basic cost-benefit analysis to weigh the financial implications of the research findings.

A. DESCRIPTIVE RESULTS

The overlaid histograms in Figure 5 illustrate the distributions of total scores assigned by computers and humans to SBIR proposals from all three prompts. The x-axis represents the total scores, marked at 25-point intervals, while the y-axis measures the frequency of proposals receiving each score. The human evaluators' scores (shown in red) are widely dispersed and have a wider variance. In contrast, the computer-generated scores (shown in blue) show a narrower distribution, characterized by a pronounced peak. Purple represents areas of overlap. Interestingly, it was challenging to produce a wider variance in computer scores that better mimics human variation. I intentionally adjusted prompt verbiage to get a wider variance but was unsuccessful. For example, Prompt 1's original structure was simply, "Evaluate the following SBIR proposals based on the following criteria and topic requirements..." followed by the rest of the content shown in the Prompt 1 text from Table 1. I then added elements to the prompt to emphasize the competitiveness of the selection process, and that only higher quality proposals could be assigned higher scores. The original structure of Prompt 1 was then modified to include:

> Critically evaluate the SBIR proposals with the understanding that only one or two will be selected for a Phase I award. This competitive context means that while each proposal may have strengths, they must be distinguished from each other with a realistic spread of scores. Not all proposals can score highly unless they stand out with exceptional qualities.

I also tried the following addition to Prompt 1, "Not all proposals can be awarded criteria A scores 35/40 or above, criteria B scores of 25/30 or above, or criteria C scores of 25/30 or above unless they demonstrate exceptional qualities." The scores mentioned in the prompt addition above consistently represented the criteria A, B, and C modal score—or the score that was given the most frequently. I was specifically trying to change the distribution of scores assigned, but the distribution remained consistent throughout the prompt refinement process.

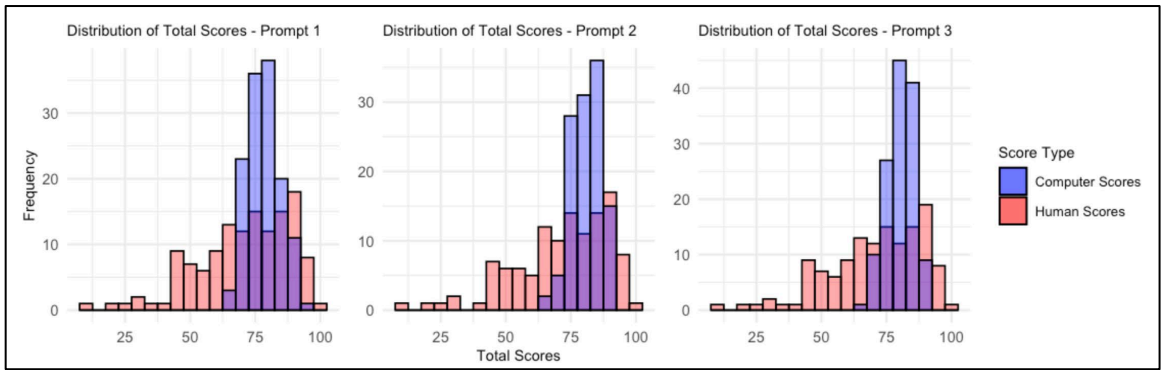


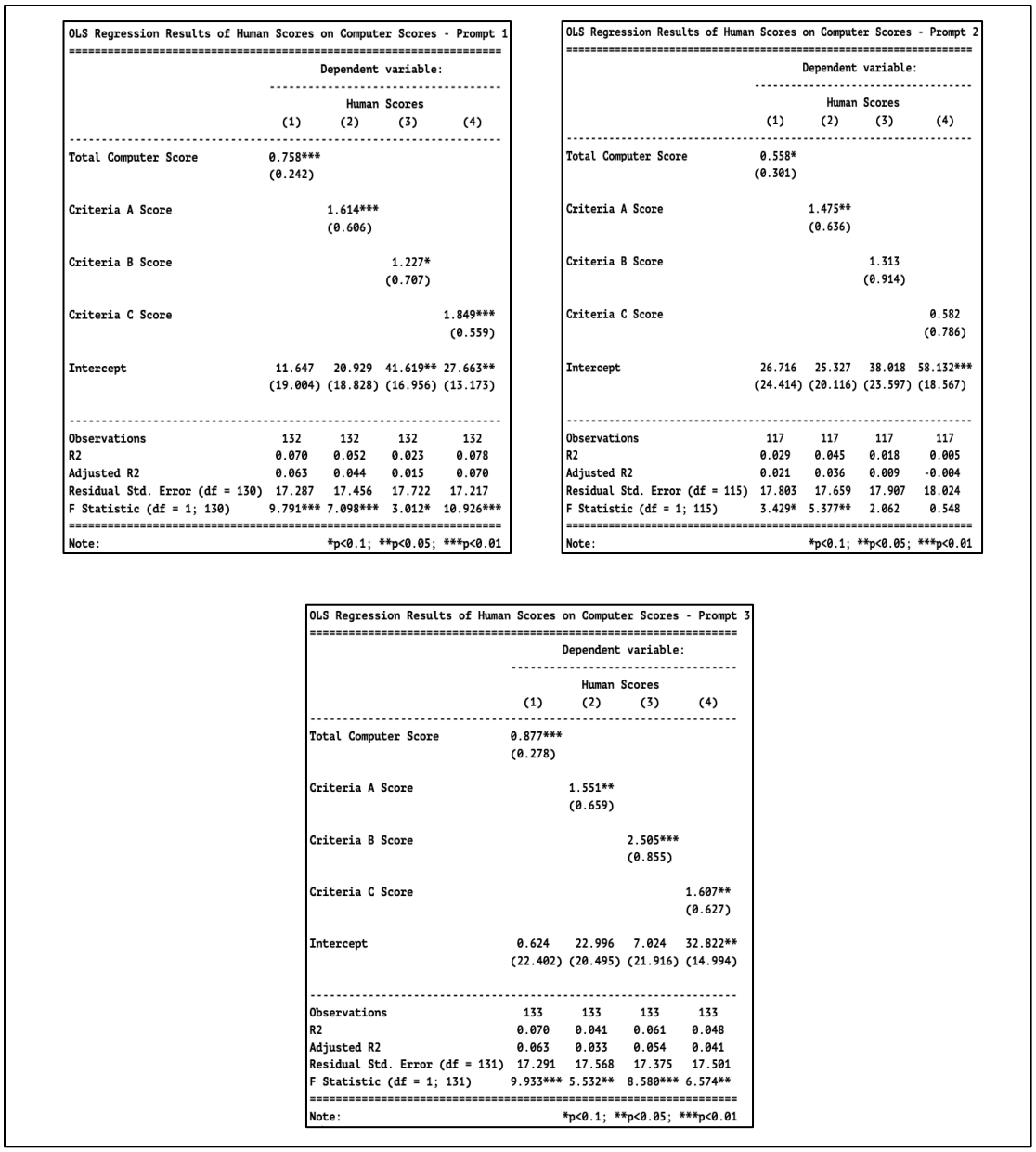
Figure 5. Prompt-Specific Distribution of Scores

This difference in scoring variation may come from several factors. For one thing, human scorers have different preferences, different scoring methodologies, and varying levels of expertise. When proposals are evaluated by humans ("Human Scores"), numerous evaluators are involved. "Computer scores" however, are produced by a single source of programming code that is constant throughout all evaluations. This difference in scoring methodologies contributes to the differences in variation. Additionally, human evaluators may be less likely to fully scrutinize proposals they initially consider to be noncompetitive, leading to lower scores that effectively remove those proposals from further consideration. Human evaluators may do this to save themselves time and to expedite their selection process. Computers already process information at a much higher rate than humans and do not consider external influences unless they're specifically programmed to do so. Therefore, computer-based solutions would not make the same convenience-based decisions that a human may be prone to make. Computer scores

would instead scrutinize each proposal equally, which may result in higher average scores and contribute to the difference in scoring distribution.

B. REGRESSION OUTCOMES – SCORING ALIGNMENT

The regression analysis captured in Figure 6 illustrates how different components of computer evaluations relate to human evaluators' total scores. Figure 6 shows the regression output from Prompts 1, 2, and 3.



OLS Regression Results of Human Scores on Computer Scores - Prompt 1

	Dependent variable:			
	Human Scores			
	(1)	(2)	(3)	(4)
Total Computer Score	0.758***			
	(0.242)			
Criteria A Score		1.614***		
		(0.606)		
Criteria B Score			1.227*	
			(0.707)	
Criteria C Score				1.849***
				(0.559)
Intercept	11.647	20.929	41.619**	27.663**
	(19.004)	(18.828)	(16.956)	(13.173)
Observations	132	132	132	132
R2	0.070	0.052	0.023	0.078
Adjusted R2	0.063	0.044	0.015	0.070
Residual Std. Error (df = 130)	17.287	17.456	17.722	17.217
F Statistic (df = 1; 130)	9.791***	7.098***	3.012*	10.926***
Note:				*p<0.1; **p<0.05; ***p<0.01

OLS Regression Results of Human Scores on Computer Scores - Prompt 2

	Dependent variable:			
	Human Scores			
	(1)	(2)	(3)	(4)
Total Computer Score	0.558*			
	(0.301)			
Criteria A Score		1.475**		
		(0.636)		
Criteria B Score			1.313	
			(0.914)	
Criteria C Score				0.582
				(0.786)
Intercept	26.716	25.327	38.018	58.132***
	(24.414)	(20.116)	(23.597)	(18.567)
Observations	117	117	117	117
R2	0.029	0.045	0.018	0.005
Adjusted R2	0.021	0.036	0.009	-0.004
Residual Std. Error (df = 115)	17.803	17.659	17.907	18.024
F Statistic (df = 1; 115)	3.429*	5.377**	2.062	0.548
Note:				*p<0.1; **p<0.05; ***p<0.01

OLS Regression Results of Human Scores on Computer Scores - Prompt 3

	Dependent variable:			
	Human Scores			
	(1)	(2)	(3)	(4)
Total Computer Score	0.877***			
	(0.278)			
Criteria A Score		1.551**		
		(0.659)		
Criteria B Score			2.505***	
			(0.855)	
Criteria C Score				1.607**
				(0.627)
Intercept	0.624	22.996	7.024	32.822**
	(22.402)	(20.495)	(21.916)	(14.994)
Observations	133	133	133	133
R2	0.070	0.041	0.061	0.048
Adjusted R2	0.063	0.033	0.054	0.041
Residual Std. Error (df = 131)	17.291	17.568	17.375	17.501
F Statistic (df = 1; 131)	9.933***	5.532**	8.580***	6.574**
Note:				*p<0.1; **p<0.05; ***p<0.01

Figure 6. Prompts 1, 2, and 3 Regression Outputs

Figure 6 shows each prompt and its respective regression output. Each regression output assesses four distinct models that use a single predictor from the computer evaluations to understand their distinct impacts on the human scores. The univariate models can all be represented in the following mathematical framework:

Human Scores $= \beta_0 + \beta 1$*(total computer score or criteria A, B, C scores) + ϵ

Since the models are all similarly constructed and predicting the same outcome, the models can all be interpreted as described in the methodology section. The predicted "Human Scores" value is equal to the sum of the intercept (β_0), the $\beta 1$ coefficient multiplied by the total computer score or criteria-specific score (depending on model selection), and the error term. An important statistical output from these models is the R^2 value (shown R2 in the figures). R^2 represents the "explainability" of the model, ranging from 0–1. The higher the score, the better the model is at representing its true relationship with the predicted variable. In this context, R^2 represents how well a linear function of the computer score explains the variation in "Human Scores." The R^2 values range from 0.005 to 0.078, suggesting that while the models have predictive power, most of the variability in human scores is not captured by these computer evaluation components. A closer look at Figure 6 shows varying degrees of statistical significance, which is indicated by the presence of asterisk marks. The number of asterisks signifies the degree of significance, ranging from 90%, 95%, to 99% confidence levels, corresponding respectively to one, two, or three asterisks. With that understanding of statistical significance, Figure 6 shows that Prompts 1 and 3 are statistically more correlated with human scores than Prompt 2 is, suggesting that the flipped interaction prompt (Prompt 2) may not be as suitable for proposal evaluation as the other prompt patterns.

The interpretation from Figure 6 is reinforced by the binned scatter plot in Figure 7, which graphically represents the mathematical relationships derived from Figure 6. The x-axis of Figure 7 categorizes the computer scores into deciles, while the y-axis displays the corresponding average human scores. If human and computer scores align well, we would expect to see low computer decile scores align with low average human scores. We would expect that progressively increasing computer decile scores should align with progressively increasing human scores. The degree of slope of the line—how steep the slope is—combined with the proximity of the data points relative to the line visually correspond with prompt performance. Of note, the top decile computer scores for all three prompts had average human scores of 77, 76, and 74.

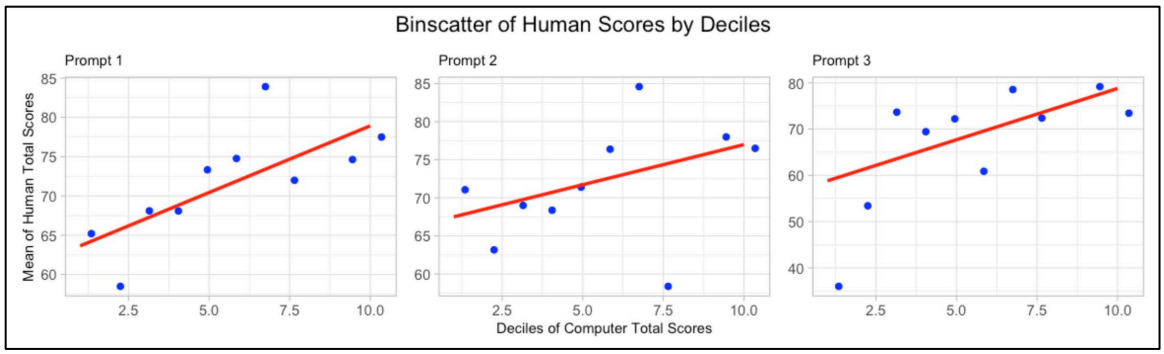


Figure 7. Binscatter of Human Scores by Deciles

Figure 7 only shows a slightly steeper slope in Prompt 1, while Prompts 2 and 3 have visually similar slopes and are better differentiated from information in Figure 6. The combined interpretations from Figure 6 and Figure 7 indicate that human and computer scoring do generally align, with modest differences in performance based on prompt selection.

C. CLASSIFICATION AND COMPETITIVENESS

After analyzing the overall alignment of scores, I then wanted to assess how well the computer evaluations could identify noncompetitive proposals. The goal was to see if computer-generated evaluations could filter out weaker proposals, which would reduce the workload for human evaluators. This is a way to frame the problem as a binary classification task. The first step was to visually assess the relationship between computer-generated scoring and human classification of competitiveness, shown in Figure 8.

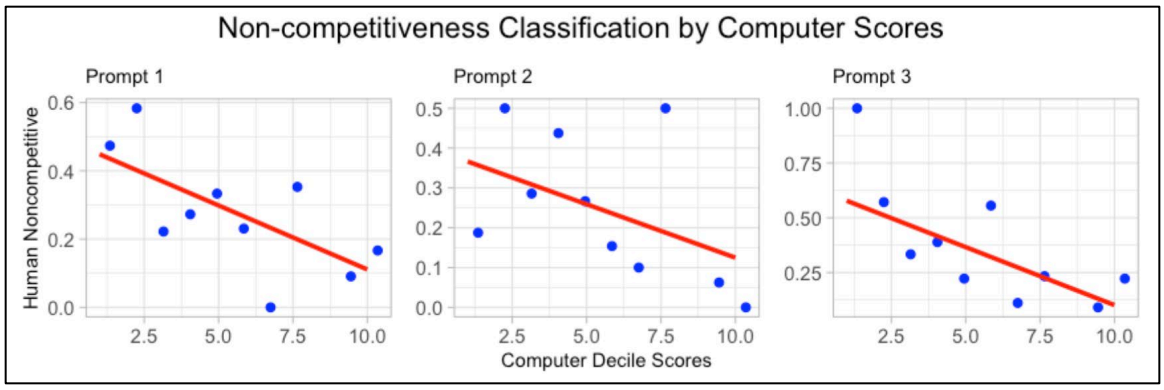


Figure 8. Noncompetitiveness Classification by Computer Scores

From Figure 8, we can see that all three prompts returned a negative sloping line. This negative slope represents the likelihood of being classified as noncompetitive as computer decile scores increases. This relationship makes sense intuitively, because we should expect to see classification as noncompetitive decrease as the score increases. Here we see computer decile scores in the ninth and tenth deciles tend to have around a 10% to 20% probability of being classified as noncompetitive, with a general increase in probability as decile score decreases. As previously stated in Chapter 3, there is no point at which a human evaluation team would formally determine or assign a proposal as noncompetitive.

I needed to turn this classification of noncompetitive vs. competitive challenge into a binary measure so that I could use standard machine learning accuracy metrics. I chose to classify bottom quartile scores as noncompetitive. Making this indicator requires some explanation because the proposals differ by group or topic. Since each topic (group) of proposals differs in terms of its own requirements and evaluation team, I chose to compute the average bottom quartile score as a threshold for classification. This methodology was used for both for the human and computer scores so that each had a normalized classification threshold that was specific to their respective average bottom quartile scores. Normalizing in this way also helped to mitigate the impact of the variance in scores between human and computer scores.

The next step in this process was to validate the computed threshold score or try to visually identify the optimal threshold score for classification. The histograms in

Figure 9 show the prompt-specific options for score thresholds. The y-axis represents the frequency of proposals, while the x-axis is the computer's assigned total score. The proposals are then divided by human classification as competitive and noncompetitive, shown in green and red, respectively. The dashed vertical blue lines represent the computed thresholds from the average bottom quartile scores as described in Chapter III. In general, the competitive proposals have higher computer scores than the noncompetitive ones. However, there is considerable overlap in each prompt, which makes classification efforts more challenging. Ideally, there would be a distinct separation between competitive and noncompetitive scores. In this case, the threshold is not as clear. Selecting a higher threshold would correctly classify a higher proportion of truly noncompetitive proposals but would also misclassify a higher proportion of competitive proposals as noncompetitive.

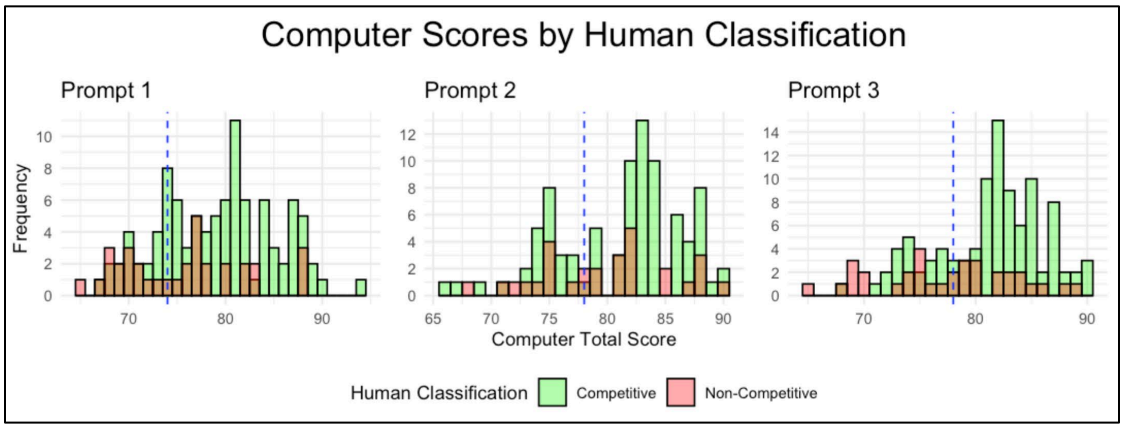


Figure 9. Computer Scores by Human Classification

D. CONFUSION MATRICES COMPARISON

A confusion matrix is a way to visualize the overall performance of a model at a given threshold. Confusion matrices show the proportion of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For this analysis, the human classification was considered the truth and the computer classification was the prediction. In this case, truly noncompetitive proposals (as classified by humans) that the computer also classifies as noncompetitive would represent TPs. Human classified

competitive proposals that the computer predicts to be noncompetitive would represent the FPs. Figure 10 illustrates the performance of the three different prompts.

Confusion Matrices			
		Actual	
		Class 0	Class 1
Prediction	Class 0	True Negative (TN)	False Negative (FN)
	Class 1	False Positive (FN)	True Positive (TP)

Prompt 1 - 71% Accuracy			
		Actual	
		Competitive	Non-Competitive
Prediction	Competitive	79	19
	Non-Competitive	19	15

Prompt 2 - 68% Accuracy			
		Actual	
		Competitive	Non-Competitive
Prediction	Competitive	70	18
	Non-Competitive	20	9

Prompt 3 - 72% Accuracy			
		Actual	
		Competitive	Non-Competitive
Prediction	Competitive	82	17
	Non-Competitive	20	14

Figure 10. Confusion Matrices

None of the models perform exceptionally well in terms of overall accuracy, ranging from 68% to 72% accurate. In fact, simply predicting the majority class (competitive) every time using a ZeroR method returns higher accuracy rates at 74% to 75%. The ZeroR method is a simple approach that works by basing all predictions off the majority class, ignoring all other predictor variables.

The Prompt 1 confusion matrix contains 132 total proposals. Out of 34 truly noncompetitive proposals, 15 were correctly classified as noncompetitive – representing TP's. There were 19 FN's, where the proposals were noncompetitive but were incorrectly classified as competitive. Similarly, there were 79 truly competitive proposals that were correctly classified (TN's) and 19 competitive proposals that were incorrectly classified as noncompetitive (FP's)). The accuracy is then given by the following equation (TP + TN) / (TP + TN + FP + FN). This logical walkthrough is the basis for understanding the

confusion matrices for the other two prompts, as their layout and formulaic interpretations are the same.

The distribution of competitive and noncompetitive scores is inherently imbalanced in this case, based on how noncompetitive is defined and the limited observations in this study. This heavily influences standard machine learning accuracy metrics that generally have much higher numbers of observations. Since the threshold for noncompetitive classification is the average 25th percentile, that means that every proposal above that score (which is approximately 75% of proposals, and approximately 100 out of 133 proposals in this case) will be considered competitive by default. This is important to consider when directly comparing the performance of ZeroR and the other models in this context. This comparison suggests that these models are not providing additional predictive power beyond what is achieved by simply guessing the majority class.

In all three models there were 19–20 false positives—proposals that were competitive that the models incorrectly classified as noncompetitive. This is not ideal and could undermine the value of this classification effort. Saving human evaluators time by filtering out noncompetitive proposals would progressively increase value over time but would also prevent strong proposals from award consideration. Given this trade-off, the current confusion matrix comparison is not optimal. The following section, "ROC Curve Analysis," will provide a more comprehensive view of potential thresholds, illustrating the trade-offs between true positive rates and varying levels of false positive acceptance.

E. ROC CURVE ANALYSIS

The ROC curve plot in Figure 11 presents all possible trade-offs between true and false positives for the three models. The y-axis represents the sensitivity or the true positive rate, while the x-axis represents the specificity or the false positive rate (often represented as 1—specificity in ROC curves). The origin represents a model that always predicts the 0 class, and the top right position at (X,Y) coordinates (1,1) represents a model that always predicts the 1 class—neither of which is practically useful outside of simply validating the performance of more complex models. A model that always

predicts 0 or always predicts 1 generally doesn't provide any value in its prediction, regardless of the accuracy measure. For example, in the case of predicting a rare disease diagnosis, a naïve model that was only interested in accuracy would simply always predict the 0 class, suggesting that the patient was negative for the rare disease. The fact that the hypothetical disease is known to be rare would guarantee a high level of accuracy, regardless of the patient's actual medical condition. Returning to Figure 11, these two extremes do allow us to depict all the varying threshold levels in between. The diagonal black dotted line that extends from the origin to the top right corner represents a random guess model, where the sensitivity equals the specificity.

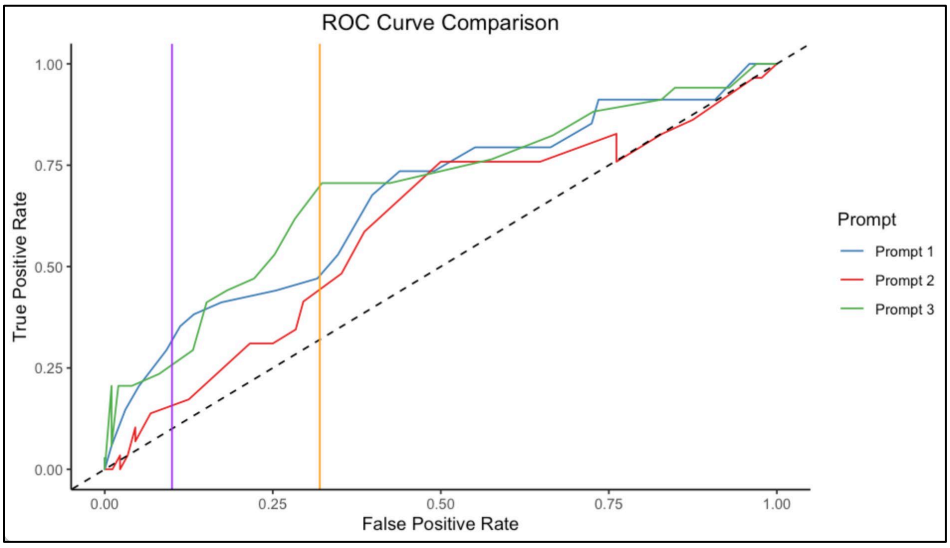


Figure 11. ROC Curve Comparison

Even through this comprehensive analysis, there is no clear optimal prompt. As a reminder, Prompt 1 was a custom prompt; Prompt 2 incorporated the flipped interaction prompt strategy; and Prompt 3 incorporated an adopted persona prompt strategy from the literature. Prompts 1 and 3 visually appear to perform better overall, and the computed area under the curve (AUC) values reinforce this observation, shown in Figure 12 below. However, statistical tests comparing the ROC curves show that there are no statistically significant differences in classifier performance. This was determined by calculating probability values (p-values): Prompt 1 vs. Prompt 2 returned p=0.377; Prompt 1 vs.

Prompt 3 resulted in p=0.769; and Prompt 2 vs. Prompt 3 had p=0.245. None of these p-values fall below the commonly used threshold of 0.05 for statistical significance. This means that there isn't enough evidence to confidently say that one model performs better than another. The vertical solid purple line in Figure 11 represents a 10% false positive acceptance level. At this acceptance level, Prompt 1 returns an approximate 30% true positive classification accuracy. At a 30% false positive acceptance represented by the orange vertical line (see Figure 11), Prompt 3 returns an approximate 70% true positive rate. These findings generally align with the confusion matrix comparisons above.

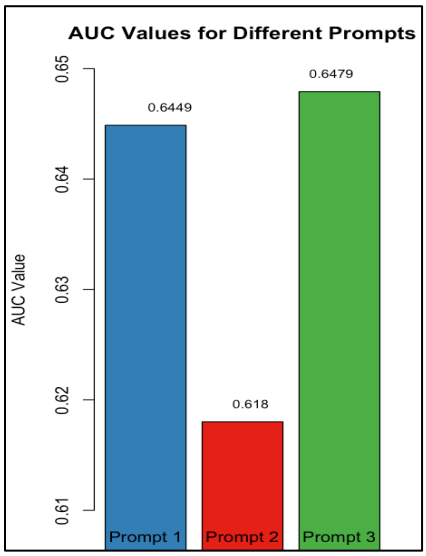


Figure 12. Area under the Curve for the Different Prompts

AUC values in the low to mid 60s suggest that the models perform slightly better than random but are not entirely reliable. There is not a clear convention on what AUC values are considered acceptable, and it will depend on the context. Examples of classification tasks that are generally considered ideal use cases are spam filtering and medical imaging classification. In these examples, the AUC values are much higher. For example, Victor Prieto et al. (2013) published a paper titled *Detecting Linkedin Spammers and its Spam Nets* that compared various algorithms and their performances in correctly classifying emails as spam. In this work, Prieto et al. found that the more robust algorithms—K-Nearest Neighbor, Decision Trees, and Naïve Bayes—had AUC values

ranging from 0.934 to 0.984, while the underperforming Support Vector Machine (SVM) achieved an AUC of 0.629. Regarding the SVM, the authors stated, "SVM achieves the worst result. ... Although this classifier has obtained a good precision and recall, it is not reliable." Similarly, Zhuoning Yuan et al. (2021) go into detail about optimizing machine learning algorithms for chest X-ray classification and melanoma detection. In their efforts, they show their robust algorithms have the leading AUC values compared to the top benchmarks on the "Chexpert" competition that Standford hosts, where machine learning algorithms attempt to assess high quality X-rays and detect chest and lung diseases. The top AUCs in this competition range from 0.906 to 0.93. This evidence suggests that to achieve dependable outcomes, especially in domains where accuracy is important, AUC values should ideally exceed the 0.90 mark. Models meeting or exceeding this threshold are considered robust and are more likely to gain acceptance in professional practice.

F. COST–BENEFIT ANALYSIS

The Defense Acquisition University (DAU) defines cost benefit analysis (CBA) at the DAU Glossary (n.d.). CBA is defined as, "An analytic technique that compares the costs and benefits of investments, programs, or policy actions in order to determine which alternative or alternatives maximize net profits. Net benefits of an alternative are determined by subtracting the present value of costs from the present value of benefits." Given that the DOD consistently has more requirements than funding, an analyst will likely be required to conduct a comprehensive CBA before implementing any LLM supported source selection tool. I will incorporate findings from this study and make assumptions necessary to conduct a preliminary CBA:

- Annual Volume: 1,000 SBIR proposals are received per year.
- Cost of Human Evaluation: The estimated cost for a typical evaluation team—which is comprised of an officer in charge, a staff non-commissioned officer in charge, and four contracted subject matter experts—to evaluate all 1,000 proposals over the course of one year is

$600,000. This assumes a combined monthly income of $50,000 for the team. $50,000 per month * 12 months per year = $600,000.

- Value of a Competitive Proposal: A Phase I award is valued at $100,000. Approximately 5% of competitive proposals can be awarded a Phase I contract. Therefore, the expected value of a competitive proposal then is 100,000 * .05 = $5,000.

- Automated Evaluation: An automated evaluation system would screen out the bottom 25% of proposals, equating to 250 proposals being removed from human consideration.

- Misclassification Rate: Based on the confusion matrix analysis, there's an assumed misclassification rate of 15% of the screened proposals. This is derived from the observed rate of (20/133) *100 = 15%. 15% of 250 proposals equates to approximately 38 proposals being misclassified.

- Benefit (Cost Saved by Automation): By filtering out 250 proposals, the automated system reduces the human evaluators' workload by 25%. The benefit in terms of cost savings is 25% of $600,000, amounting to $150,000.

- Cost (Potential Loss from Misclassification): Potentially misclassifying 38 proposals at an assumed value of $5,000 each results in a potential loss of 38 * $5,000 = $190,000.

This preliminary CBA indicates an overall present value of benefits of $150,000 saved by optimizing the efforts of human evaluators, compared to a present value of costs at $190,000 due to misclassification of competitive proposals. The net result would be an annual $40,000 loss, meaning that the automation process described in this analysis would not be recommended for immediate application. While there are still opportunities for LLMs to be used to augment acquisitions processes, this reinforces the importance of human evaluators in the proposal selection process.

V. CONCLUSIONS AND RECOMMENDATIONS

The final chapter summarizes the research findings and revisits the research questions. It acknowledges the study's limitations and presents areas for improvement. The chapter concludes by proposing areas for future research and offers actionable recommendations based on the study's insights.

A. SUMMARY OF FINDINGS

This study found a statistically significant correlation between human and computer scores. However, after using the automated scoring models as a tool for proposal classification, the study found that all three of the prompts only marginally outperformed random guessing. This suggests that while LLMs are an emerging technology with widely recognized potential, their current accuracy and reliability measures prevent them from immediately augmenting the source selection process.

B. RESEARCH LIMITATIONS

The major limitations to this study were the number of proposals included, the number of prompt engineering strategies used, and personal costs incurred. Increasing the number of prompt engineering patterns analyzed or increasing the number of proposals for the analysis would have improved the robustness of the study but would have come at an increased personal cost. As an individual, non-funded research project, the combination of token input/output costs became increasingly prohibitive throughout the conduct of this analysis.

C. RESEARCH QUESTIONS

1. How does the performance of automated evaluations using LLMs compare to that of human evaluators in the SBIR proposal evaluation process?

The regression analysis showed a general alignment between automated and human evaluations. However, the effectiveness varied based on the specific prompt used.

This area of study would benefit greatly by including more proposals for overall robustness and assessing the effectiveness of a variety of prompts engineering strategies.

2. To what extent can automated evaluations using LLMs classify competitive and noncompetitive proposals?

The LLMs demonstrated a modest ability to classify proposals into competitive and noncompetitive categories, with AUC values indicating a performance slightly better than random chance. Since the DOD requires performance much better than "slightly better than random chance," further research is required to assess possible options for improving classification accuracy.

3. What are the potential broader implications of using automated evaluations using LLMs for optimizing source selection for contracts within the DOD acquisition effects beyond the SBIR program?

This analysis suggests that LLMs have promising potential to streamline the source selection process and have application beyond the SBIR program. However, the current limitations in classification accuracy and reliability prevent the immediate integration of these technologies in broader DOD acquisition contexts. If accuracy metrics can reach a level that is similar to human evaluators, the potential benefits realized in terms of efficiency and speed would be highly advantageous to the DOD.

4. What are the challenges and limitations associated with using an LLM in the SBIR proposal evaluation process, and how can they be addressed?

The primary challenges of using an LLM in the source selection process are the costs associated with LLM usage, the current accuracy and reliability of their outputs, and how sensitive those outputs may be to different prompt engineering strategies. While the costs of LLM usage are far more easily absorbed by the DOD compared to an individual, more research into prompt engineering strategies would be required to try and maximize accuracy and reliability. Additionally, at the time of this work OpenAI has arguably the most robust capability in the GPT-4 model. However, continued

development by OpenAI and their competition on advancing their products would require re-evaluation consistent with market developments. While the GPT-4 model is currently OpenAI's most capable product, OpenAI is already beginning to advertise the release of GPT-5 for later this year. Additionally, major technology firms such as Google, Microsoft, Amazon, along with emerging startups like Anthropic and Mistral AI, are making significant strides. These entities are developing their own models that are rapidly becoming more competitive with OpenAI.

D. RECOMMENDATIONS FOR FUTURE RESEARCH

This study establishes a foundation for future analysis to easily build on. A straightforward way to build on this work would be to follow the same steps but include more proposals, or explore additional prompt engineering strategies. The programming code in R has been optimized to interact with the LLM, reference the topic requirements, loop through individual proposals within the topic, and then move on to the next topic, continuing the loop until completion. The trial-and-error process of creating this complex code took a lot of time and incurred additional costs. The personal costs associated with this coding trial-and-error has largely been absorbed by this study and therefore should not be as impactful in future studies—adding additional proposals to the existing code will run successfully with very few changes.

It is recommended that future studies further explore different prompt engineering patterns. This study included two academically sourced prompt patterns and a custom prompt due to cost and time constraints, but future studies on automated SBIR evaluations would benefit from integrating emerging concepts from the field of prompt engineering. The literature referenced in this work provides a ready-made catalog of 16 different prompt strategies that will likely have varying levels success.

Future research could also conduct this same study while varying the LLM selection. This study chose to use OpenAI based on current literature and performance metrics, but future studies could compare results from this study to LLM options provided by OpenAI's competitors such as Google, Microsoft, Amazon, etc. Such a study

would likely be of interest to the DOD, as it would essentially serve as free market research for the acquisition community.

Additional recommendations for study include classifying SBIR topics by type to assess the relationship of prompt performance relative to topic type. For example, this study included 10 different topics that could be subdivided into three categories. The categories recommended for the topics in this study are medical technology, communication/information systems, and general military equipment enhancements. Additional categories could be created as needed based on additional topic details. The intent of such a study could be to assess how well medical or communication-based topics align with human evaluations when compared to military equipment enhancements.

LIST OF REFERENCES

Audretsch, D. B., Link, A. N., & Scott, J. T. (2019). Public/private technology partnerships: Evaluating SBIR-supported research. In *The Social Value of New Technology* (pp. 264–278). Edward Elgar Publishing. https://www.elgaronline.com/display/edcoll/9781788116329/9781788116329.00021.xml

Bollinger, A. S., & Smith, R. D. (2001). Managing organizational knowledge as a strategic asset. *Journal of Knowledge Management*, *5*(1), 8–18. https://doi.org/10.1108/13673270110384365

Defense Acquisition University. (n.d.). Glossary Term. Retrieved 2 April 2024 from https://www.dau.edu/glossary/cost-benefit-analysis

Department of Defense. (n.d.). SBIR/STTR Program. Retrieved March 21, 2024 from https://www.defensesbirsttr.mil/SBIR-STTR/Program/

Department of Defense. (n.d.). SBIR/STTR Guide. Retrieved March 21, 2024, from https://www.defensesbirsttr.mil/SBIR-STTR/Guide/

Department of the Navy. (2017). Small Business Innovation Research (SBIR) Small Business Technology Transfer (STTR) Topic Author and Topic Reviewer Guidebook (Version 1.1). [PDF].

Galope, R. V. (2014). What Types of Start-ups Receive Funding from the Small Business Innovation Research (SBIR) Program?: Evidence from the Kauffman Firm Survey. *Journal of technology management & innovation*, 9(2), 17–28. https://www.scielo.cl/scielo.php?pid=S0718-27242014000200002&script=sci_arttext

George, A. S., & George, A. H. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal*, *1*(1), 9–23. https://puiij.com/index.php/research/article/view/11

Ghayvat, H., Awais, M., Bashir, A. K., Pandya, S., Zuhair, M., Rashid, M., & Nebhen, J. (2022). AI-enabled radiologist in the loop: Novel AI-based framework to augment radiologist performance for COVID-19 chest CT medical image annotation and classification from pneumonia. *Neural Computing and Applications*, 35, 14591–14609. https://doi.org/10.1007/s00521-022-07055-1

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). *How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection*. ArXiv [preprint]. http://doi.org/10.48550/arXiv.2301.07597

Held, B., Edison, T., Lawrence Pfleeger, S., Anton, P. S., & Clancy, J. (2006). *Evaluation and recommendations for improvement of the department of defense small business innovation research (SBIR) program*. RAND. https://apps.dtic.mil/sti/citations/ADA457417

Kautz, H. . (2022). The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, *43*(1), 105–125. https://doi.org/10.1002/aaai.12036

MITRE Corporation. (2016). Slow defense acquisitions cost lives. In *Opportunities for Impact* (No. 16–3837). https://www.mitre.org/sites/default/files/publications/mitre-corporation-slow-defense-acquisitions-cost-lives-october-2016.pdf

Mun, J. (2023). Management and business knowledge representation for decision making: Applying artificial intelligence, machine learning, data science, and advanced quantitative decision analytics for making better-informed decisions. (SYM-AM-23-095). Acquisition Research Program. https://dair.nps.edu/handle/123456789/4862

Nasir-Moin, M., Suriawinata, A. A., Ren, B., Liu, X., Robertson, D. J., Bagchi, S., Tomita, N., Wei, J. W., MacKenzie, T. A., Rees, J. R., & Hassanpour, S. (2021). Evaluation of an Artificial Intelligence-Augmented Digital System for Histologic Classification of Colorectal Polyps. *JAMA Network Open*, *4*(11), e2135271–e2135271. https://doi.org/10.1001/jamanetworkopen.2021.35271

OpenAI. (2023). GPT-4 technical report. *arXiv*. https://arxiv.org/abs/2303.08774v3

OpenAI. (n.d.) *Introduction*. Retrieved February 5, 2024, from https://platform.openai.com/docs/introduction

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

Prieto, V. M., Álvarez, M., & Cacheda, F. (2013). Detecting Linkedin spammers and its spam nets. *International Journal of Advanced Computer Science and Applications*, *4*(9), 189–199. Research Gate. https://www.researchgate.net/publication/314432574_Detecting_Linkedin_Spammers_and_its_Spam_Nets

Small Business Administration. (2020). *Small Business Innovation Research (SBIR) program overview*. From https://www.sbir.gov/sites/default/files/SBA_SBIR_Overview_March2020.pdf

Small Business Innovation Research/Small Business Technology Transfer. (n.d.). *About*. 2023, from https://www.sbir.gov/about

Small Business Innovation Research/Small Business Technology Transfer. (n.d.). *Method of selection and evaluation criteria*. Retrieved June 9, 2023, from https://www.sbir.gov/content/method-selection-and-evaluation-criteria

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. In *Proceedings of the 34th Neural Information Processing Systems (NeurIPS)*. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

Wallner, M., Peterson, J., Swearingen, W., Zook, M. V., & Gaster, R. (2021). SBIR: A Catalyst for Innovation in the Flyover States. *Issues in Science and Technology*, 37(3), 82–84. https://nps.primo.exlibrisgroup.com/discovery/fulldisplay?docid=cdi_proquest_journals_2530028300&context=PC&vid=01NPS_INST:01NPS&lang=en&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,SBIR

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT.* ArXiv. http://doi.org/10.48550/arXiv.2301.07597

Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2023). Large language models as optimizers. *arXiv:2309.03409v2 [cs.LG]*. Google DeepMind. https://arxiv.org/abs/2309.

Yuan, Z., Yan, Y., Sonka, M., & Yang, T. (2021). Large-scale robust deep AUC maximation: A new surrogate loss and empirical studies on medical image classification*. 2021 IEEE/CVF International Conference on Computer Vision,* pp. 3020–3029. https://www.doi.org/10.1109/ICCV48922.2021.00303

Acquisition Research Program
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net