EVALUATING SBIR PROPOSALS: A COMPARATIVE ANALYSIS USING ARTIFICIAL INTELLIGENCE AND STATISTICAL PROGRAMMING IN THE DOD ACQUISITIONS PROCESS

NAVAL POSTGRADUATE SCHOOL

Abstract

- Assessment of Large Language Models' (LLM) ability to automate classification of acquisition proposals as either competitive or noncompetitive.
- This classification aims to establish a faster, more consistent, and objective evaluation system when compared to human assessment.
- Three different prompt engineering strategies were used and compared against one another.
- Interaction with the LLM was conducted via R programming and OpenAI application programming interface—not the standard graphical user interface.

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

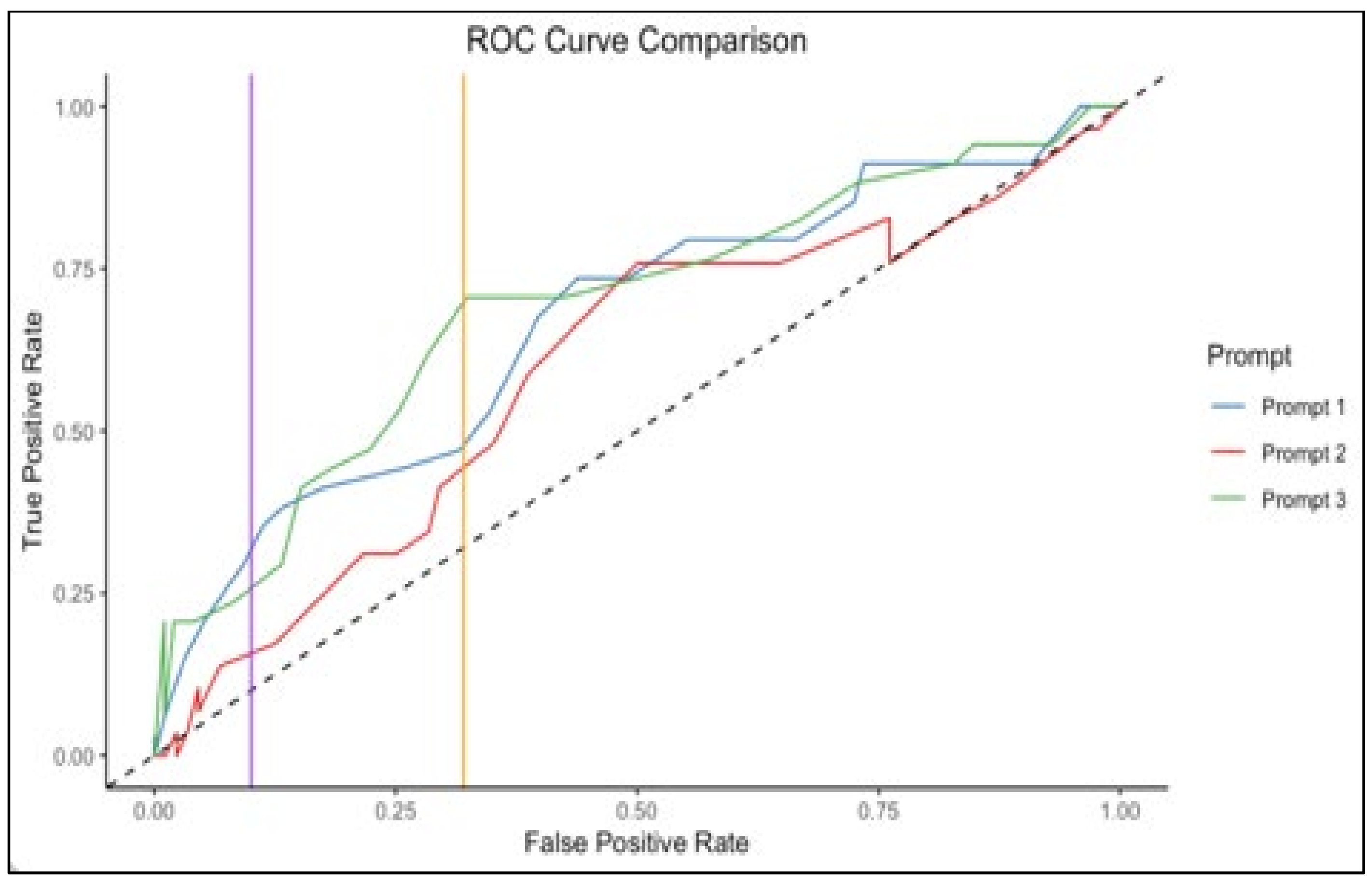
		Actual	
		Competitive	Non-Competitive
Prediction	Competitive	70	18
	Non-Competitive	20	9

Prompt 3 - 72% Accuracy

		Actual	
		Competitive	Non-Competitive
Prediction	Competitive	82	17
	Non-Competitive	20	14

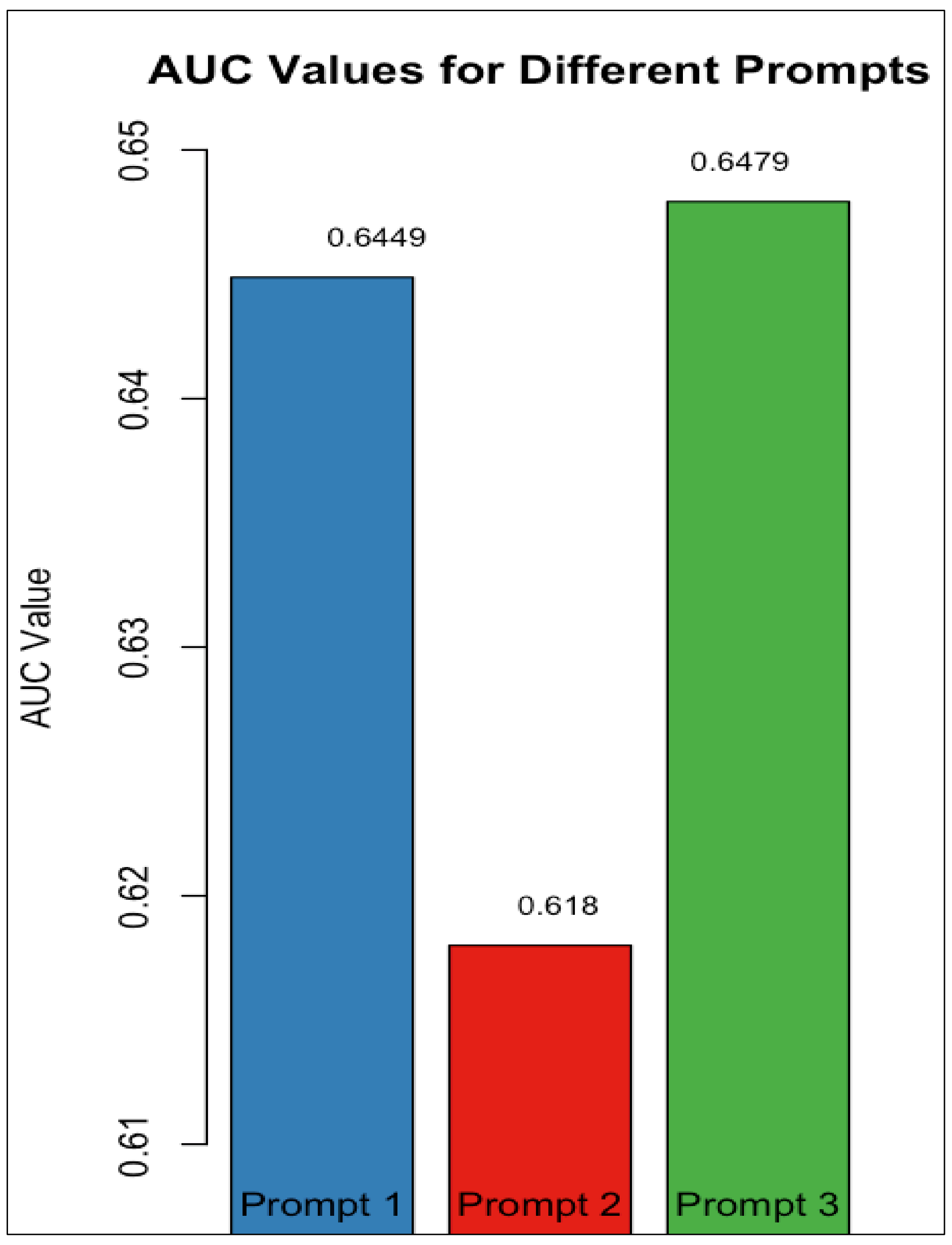
Methods

- Ordinary Least Squares Regression was used to assess the alignment of scoring between human and computer-generated scores.
- Machine Learning accuracy metrics were used to determine how well the constructed models performed in classification relative to human classification.
  - Human evaluation does not typically involve formal classification in this manner. This study used the bottom quartile score for human evaluations as the classification threshold.



Results & Their Impact

- Prompt 1 (custom prompt/AUC=0.6449) and Prompt 3 (adopted persona prompt/AUC=0.6479) appear to perform better than Prompt 2 (flipped interaction prompt/AUC=0.6180).
- AUC values in the low to mid 60's suggest that the models perform slightly better than random guess models (where AUC is equal to 0.5) but they are not entirely reliable.
- However, statistical tests comparing the ROC curves show that there is no statistically significant differences in classifier performance.
- Literature suggests that in domains where accuracy is extremely important, AUC values should ideally exceed the 0.90 threshold to be considered dependable.
- Recommend continued study using: 1) More proposals, 2) Broadened scope of prompt engineering strategies, 3) Different LLM performance comparison



Department of Defense Management
www.nps.edu/ddm

ACQUISITION RESEARCH PROGRAM
ARP

Cullen Tores, Major, USMC
Advisors: Dr. Maxim Massenkoff
Dr. Robert Mortlock