ASSESSING DOD CONFIDENCE & BIAS IN AI/LLM AUTHORED EVALUATION FACTORS

Naval Postgraduate School

Abstract

AI/LLMs have shown promise in various tasks, but their use in authoring source selection evaluation factors in the Department of Defense (DOD) is not well-studied. Understanding the effectiveness of AI-authored evaluation factors is crucial for reliable decision-making. The integration of LLM technology in the DOD aligns with the rise of AI. This exploratory analysis investigated DOD acquisition professionals' confidence in and bias toward AI-authored source selection evaluation factors. Surveys at George Mason University (GMU) and the Naval Postgraduate School presented professionals with requirements documentation and human or AI-generated evaluation factors. Quantitative analysis showed lower confidence in AI-authored factors. Qualitative feedback indicated deficiencies in clarity and traceability. Furthermore, when told factors were AI-authored, average ratings decreased, revealing slight algorithm aversion, especially among older professionals. The findings imply limitations in AI-authored evaluation factors. The authors recommended the development of an acquisitions AI guide to aid responsible use in acquisitions. Further research with larger, varied samples and various AI tools is needed. This initial work informs AI integration policies to balance innovation and public trust in defense acquisitions.

Research Question 1

Do DOD acquisition professionals have confidence in AI-authored evaluation factors?

Research Question 2

Do DOD acquisition professionals demonstrate bias when they believe they are reading AI-authored evaluation factors?

Methods

1. Collect real/current gov PWS and Section M
2. Have ChatGPT create section M using gov PWS
3. Create Survey
   I. Without knowing authorship, what does the participant think about the quality of the section M?
   II. Tell participant authorship, but only tell 50% the truth.
   III. Did knowing the author was AI or Human change the participants ratings?
4. Conduct GMU (24 people) & NPS (19 people) Surveys
5. Exploratory Analysis of survey responses
6. Recommendations provided to improve quality/perceptions of AI/LLM authored materials

Table 1. 2×2 Factorial Design Conditions for the GMU Survey

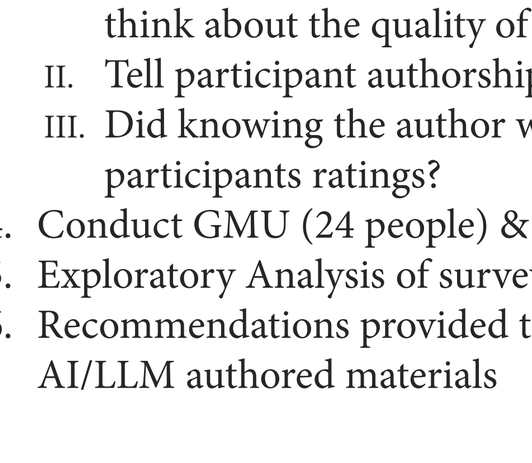
	Disclosure of Authorship: True (B1)	Disclosure of Authorship: Not True (B2)
Source of Criteria: Human-authored (A1)	Group 1: Human – Human	Group 2: Human – AI
Source of Criteria: AI-authored (A2)	Group 3: AI – AI	Group 4: AI – Human

This exploratory analysis is structured into four groups based on who the author is (Human or AI), and disclosure (Truth or Lie).

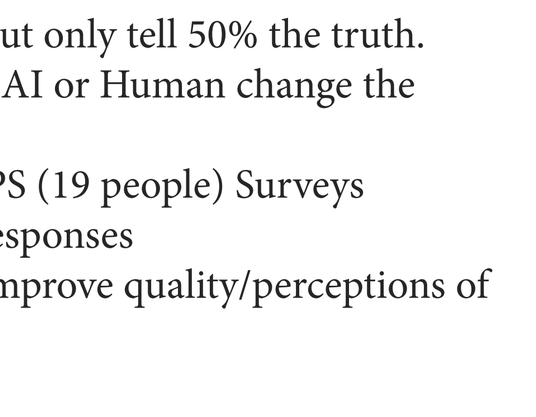
Results & Their Impact

DOD professionals must have the knowledge and skill to assess the quality of AI/LLM-authored materials. Poorly generated materials, used to support the mission, could cost the government time and money and increase the risk to people and property.

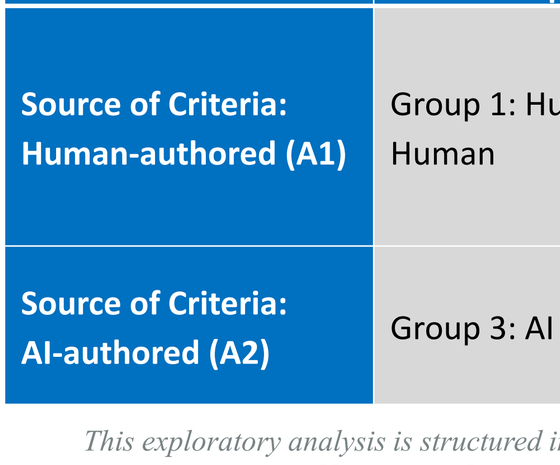
DOD professionals must exercise caution when working with technology. They must be able to recognize their own bias towards hesitation of its adoption (algorithm aversion or AI distrust) or overconfidence in its capabilities (automation bias). Delaying the adoption of technology could inhibit the mission, but so could overconfidence in its capabilities.



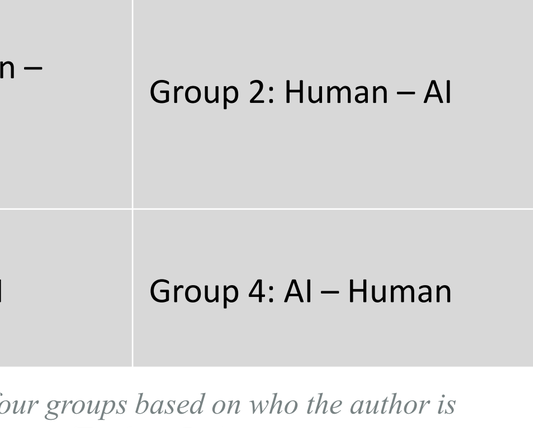
Participants' rating AI-authored factors were lower than when rating human-authored factors.



Participants have significantly less confidence in AI-authored evaluation factors.



The mean rating for participants who were told human-authorship (groups 1 and 4) increased while those who were told AI-authorship (groups 2 and 3) decreased.



Groups 1 and 4 combined (Told-Human) showed an increase in confidence, regardless of actual authorship. In contrast, groups 2 and 3 (Told-AI) decreased.

Acquisition Research Program

www.acquisitionresearch.net

S. Chad Hedgepeth, USAF
Ryan Mark D. Tagatac, Capt, USAF

Advisors: Lt Col Daniel J. Finkenstadt
Dr. Brett W. Josephson
Lt Col Jamie M. Porchia