

INTRODUCTION

Much debate and ambiguity exist surrounding the use of Artificial intelligence (AI) specifically in scientific research writing. With the speed at which AI products are developing, it is unknown whether the quality of its products has risen to the level of human authors. This study compares the quality of AI- vs-human-generated plastic surgery scientific abstracts through independent evaluation and by testing detection via side-by-side comparison across plastic surgeons, trainees, medical students, and laypersons.

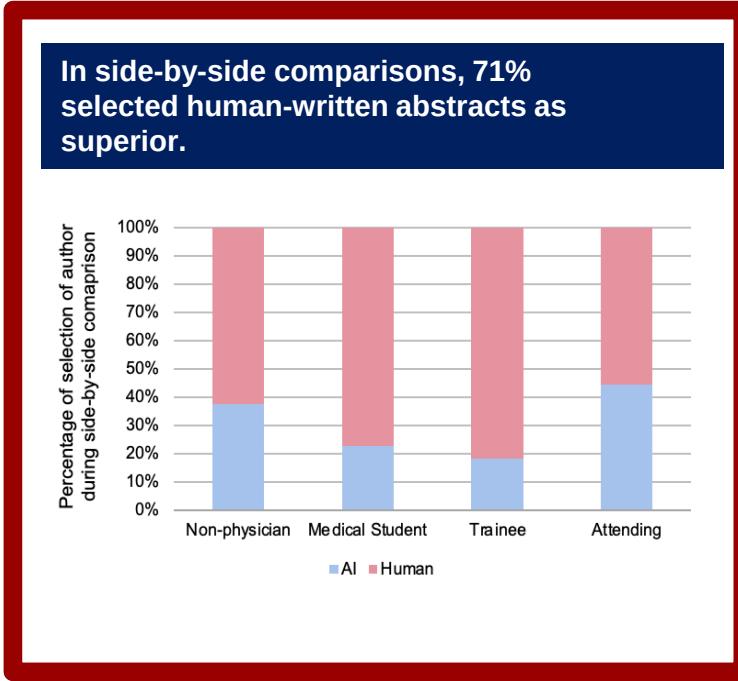
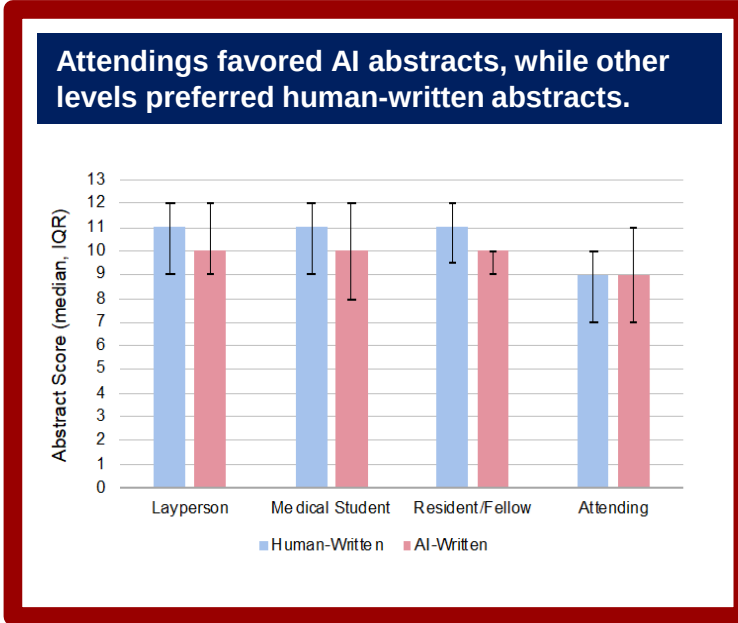
METHODS

- Ten abstracts were created using study outlines of previously accepted plastic surgery studies: 5 were generated using Pre-trained Transformer 4 (GPT-4) and 5 were written by plastic surgery trainees
- Surveys were electronically distributed to participants including plastic surgeons, trainees, and laypeople who were blinded to authorship.
- Respondents were asked to rate 8 abstracts individually using a 12-point quality scale and compared 4 abstracts side-by-side.
- Writing quality and the association between training levels and discrimination abilities were assessed using paired T-tests and regression models

RESULTS

- Of 355 surveys, 174 (49%) were completed. Respondents averaged 29 years, 56% were females, 68.5% Caucasians, 51% hold/pursue doctorate degrees, and 27.3% had a bachelor's degree or lower.
- Respondents in medicine constituted 63% medical students, 21% residents, and 17% attendings. Mean scores showed no significant difference between AI and human-written abstracts (9.9 vs. 10.2, P=0.11).
- Risk-adjusted analysis showed medical students and residents more likely to prefer human-written abstracts over AI, compared to laypeople and attendings (P<0.001).

Domain	Criteria Response	Human-written N = 202	AI-written N = 201	P-value
Methods	Meets criteria	146 (72.3%)	142 (71.0%)	0.58
	Partially meets criteria	50 (24.8%)	48 (24.0%)	
	Does not meet criteria	6 (3.0%)	10 (5.0%)	
Results	Meets criteria	128 (63.4%)	129 (65.2%)	0.76
	Partially meets criteria	62 (30.7%)	55 (27.8%)	
	Does not meet criteria	12 (5.9%)	14 (7.1%)	
Discussion	Meets criteria	112 (56.0%)	95 (47.5%)	0.19
	Partially meets criteria	66 (33.0%)	83 (41.5%)	
	Does not meet criteria	22 (11.0%)	22 (11.0%)	
Writing Quality	Meets criteria	120 (59.4%)	97 (48.3%)	0.053
	Partially meets criteria	69 (34.2%)	82 (40.8%)	
	Does not meet criteria	13 (6.4%)	22 (10.9%)	



CONCLUSIONS

- GPT-4 can generate scientific abstracts comparable to human-written.
- Trainees exhibit better differentiation abilities, and all readers discriminate better when compared side-by-side.
- The implications of AI in scientific writing warrant discussion due to its human-level content synthesis capability.

DISCLOSURES

The opinions expressed are not the official position of the United States Government, United States Army or Department of Defense.

REFERENCES

- Sallam M. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023;11(6). doi:10.3390/healthcare11060887
- Mackenzie EL, Larson JD, Poore SO. Plastic surgery and specialty creep: an analysis of publication trends. *Arch Plast Surg*. 2021;48(6):651-659. doi:10.5999/aps.2021.00745
- Temsah O, Khan SA, Chaiah Y, et al. Overview of early chatgpt's presence in medical literature: insights from a hybrid literature review by chatgpt and human experts. *Cureus*. 2023;15(4):e37281. doi:10.7759/cureus.37281
- Spake CSL, Zeyl VG, Crozier JW, Rao V, Kallinen LK. An analysis of publication trajectory in plastic surgery across the decades. *J Plast Reconstr Aesthet Surg*. 2022;75(1):439-488. doi:10.1016/j.jbps.2021.09.039

 @pennplasticsurgery  @PennPlasticSurg