SAEM 2026 · EDUCATION / AI

Artificial Intelligence and the Evolution of Residency Personal Statements: A Natural Experiment in Emergency Medicine

R. Andrew Taylor, J. Swap, Ferial Delavari, F. Turkistani, Joshua Easter, Winston Wu

Department of Emergency Medicine, University of Virginia · UVA Health

AUC 0.50–0.76

AI-DETECTOR PERFORMANCE RANGE ACROSS TOOLS

From chance-level to moderate · 0.5–29% sensitivity at 1–5% false-positive rate

4,166
EM RESIDENCY PERSONAL STATEMENTS ANALYZED
3,404 pre-AI (2018–2023) · 762 post-AI (2024)

800
SYNTHETIC AI COHORT
Generated for benchmarking

d = −0.314
TYPE-TOKEN RATIO SHIFT POST-AI
0.4988 → 0.5097 · p < 0.001

BACKGROUND

Are personal statements still authentic in the AI era?

Generative AI tools may alter the authenticity and evaluative utility of residency personal statements.

Aim. Determine whether EM residency personal statements changed in the generative AI era, and assess the accuracy and false-positive risk of common AI-detection tools.

METHODS

Retrospective de-identified analysis with synthetic AI cohort.

Design. Retrospective analysis of de-identified EM residency personal statements submitted from 2018–2024.

Cohorts. Pre-AI human (2018–2023) · post-AI era (2024) · synthetic AI cohort (n=800).

Features. Linguistic, stylometric, semantic — robust lexical diversity (MTLD/HD-D), syntactic variability, repetition/formulaicity indices, embedding-based semantic dispersion, between-statement similarity.

Statistical tests. Mann–Whitney U · Cohen's d · FDR control for secondary features.

Detector benchmarking. AUC with 95% CIs and false-positive rates across pre, post, and synthetic statements.

RESULTS

Convergence post-AI; detectors inconsistent.

4,166	Total personal statements analyzed (3,404 pre, 762 post)
d = −0.266	MTLD pre vs post-AI (0.8210 → 0.8262) · p < 0.001
d = −0.314	Type-Token Ratio (0.4988 → 0.5097) · p < 0.001
d = −0.314	Distinct unigrams · p < 0.001
d = 0.218	Word repetition index decrease · p < 0.001
d = 0.267	Burstiness decrease · p < 0.001
d = −0.050	Mean pairwise similarity (0.4705 → 0.4777) · p < 0.001
Convergent	Tighter narrative clustering (covariance trace 0.5295 → 0.5223)
AUC ~0.76	Best detector · 13–29% sensitivity at 1–5% FPR
AUC ~0.50	Worst detector · chance-level (0.5–1.5% sensitivity at 1–5% FPR)

CONCLUSION

Detectors are inconsistent — caution warranted in admissions screening.

EM residency personal statements show measurable convergence in the post–generative-AI era. Current AI-detection tools exhibit inconsistent performance and may yield nontrivial false-positive rates in real-world application materials, raising concerns about their use in high-stakes admissions screening.

ASK ME ABOUT

- Stylometric feature design
- Synthetic cohort construction
- Detector benchmarking
- Admissions-screening implications