

Stanford's Detector-Bias Audit — GPT Detectors Flag Non-Native English Essays as AI-Written

AI in Education · Good practice · United States of America

Quality score: 60/100

Evidence strength: Moderate

Summary

Stanford researchers found seven GPT detectors misclassified over half of 91 human-written TOEFL essays as AI text while judging US 8th-grade essays near-perfectly, a warning for any school using detectors.

Key dimensions

Dimension	Score
Evidence of learning impact	1/3
Equity & inclusion	3/3
Transparency, ethics & data protection	3/3
Teacher capability & pedagogy	1/3
Scalability & sustainability	1/3

Evaluated by C-NAPSE

Data sources

[C-NAPSE web research](#) — accessed 2026-10-01

[Stanford University \(Department of Biomedical Data Science\)](#) — accessed 2026-10-01

[Patterns paper \(Cell Press\)](#) — accessed 2026-10-01

[ScienceDaily summary](#) — accessed 2026-10-01

Cite this — Evidoria. *Stanford's Detector-Bias Audit — GPT Detectors Flag Non-Native English Essays as AI-Written*. Persistent ID: a6197152-aa56-4907-9995-196fb61181ae.