Grading at Scale — A Randomized Trial of AI vs. Human Graders Across Four Political-Science Programs

AI in Education · Good practice · United States of America

Quality score: 67/100

Evidence strength: Randomised controlled trial

Summary

A peer-reviewed RCT across four universities randomly assigned 3,080 short-answer responses to GPT-4 or human grading. Scores were statistically indistinguishable, and AI let instructors give large classes feedback matching classes a third the size.

Key dimensions

Dimension	Score
Evidence of learning impact	1/3
Equity & inclusion	1/3
Transparency, ethics & data protection	2/3
Teacher capability & pedagogy	3/3
Scalability & sustainability	3/3

Evaluated by C-NAPSE

Data sources

C-NAPSE web research — accessed 2026-08-20

University of Houston, with University of South Carolina, University of Michigan & Academia Sinica (Taiwan) — accessed 2026-08-20

PMC full text (Heinrich et al., PLoS One 2025) — accessed 2026-08-20

Semantic Scholar record — accessed 2026-08-20

Cite this — Evidoria. *Grading at Scale — A Randomized Trial of AI vs. Human Graders Across Four Political-Science Programs*. Persistent ID: 49bab669-e535-477b-b874-00b96311cc56.