World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

The evaluator effect: Reliability challenges in AI-assisted assessment for business and educational technology

Francis Melaragni * and Phyllis Baron

MCPHS University, 179 Longwood Ave. Boston, MA 02115

World Journal of Advanced Engineering Technology and Sciences, 2026, 18(03), 497-514

Publication history: Received on 12 February 2026; revised on 25 March 2026; accepted on 27 March 2026

Article DOI: https://doi.org/10.30574/wjaets.2026.18.3.0182

Abstract

As educational institutions increasingly adopt large language models (LLMs) for automated assessment, grading support, and instructional feedback, fundamental questions emerge about the reliability of these systems. This study investigates whether LLM evaluators reach consistent conclusions when assessing identical student-like outputs—a critical validity concern for evidence-based educational practice. Five leading LLMs generated responses to seven domain-diverse prompts simulating typical educational tasks. Two LLM evaluators (Claude Sonnet 4 and ChatGPT-4o) independently assessed each response using standardized rubric criteria common in higher education. Results revealed systematic instability: Claude Sonnet 4's performance rating varied from 22.9% to 74.3% of criterion wins depending solely on which AI system conducted the evaluation—a 51.4 percentage point swing. This evaluator-dependent variance produced near-complete rank reversals, raising serious questions about the validity of AI-assisted assessment in educational contexts. The findings have immediate implications for: (1) faculty adopting AI grading tools, (2) institutions procuring educational technology, (3) researchers using AI evaluators in learning analytics, and (4) administrators developing AI governance policies. We argue that educational technology stakeholders must implement multi-evaluator validation protocols, transparent reliability reporting, and careful pilot testing before deploying AI assessment systems. This study contributes to the growing literature on responsible AI integration in teaching and learning while highlighting both opportunities and risks for educational innovation.

Keywords: Educational Technology Assessment; AI Reliability; Automated Grading; Learning Management Systems; Educational Technology Evaluation; Pedagogical Technology; Faculty Development

1. Introduction: ai assessment tools in educational practice

1.1. The Promise of AI-Assisted Assessment in Higher Education

Educational institutions face mounting pressure to provide timely, personalized feedback while managing growing enrollments and constrained resources. AI-powered assessment tools promise to address these challenges by augmenting human grading, providing immediate feedback on student writing, and supporting formative assessment at scale (Holstein et al., 2019). Major learning management systems now integrate AI writing assistants, while standalone platforms offer automated essay scoring and feedback generation for classroom use.

The educational appeal is substantial. Faculty report spending 30-40% of instructional time on assessment activities (Gibbs & Simpson, 2004), time that could potentially be redirected toward high-impact pedagogical interactions. AI tools offer consistent rubric application, reduced grading time, and the possibility of providing detailed, immediate feedback that scaffolds student learning (Shute, 2008). These potential benefits have accelerated institutional adoption,

* Corresponding author: Francis Melaragni.

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0.

with surveys indicating that over 60% of higher education institutions are piloting or deploying AI assessment technologies (Thompson et al., 2024).

1.2. The Assessment Validity Problem

However, the rapid integration of AI assessment tools into educational practice has proceeded faster than empirical validation of their reliability and pedagogical soundness. When faculty adopt AI grading assistants or institutions procure automated assessment systems, they implicitly assume these tools provide consistent, valid evaluations aligned with learning objectives. This assumption—that competent AI evaluators using standardized criteria will reach broadly similar conclusions about student work quality—has received surprisingly limited empirical scrutiny in educational contexts.

The magnitude of potential consequences extends across multiple educational stakeholders. For students, unreliable AI assessment could mean receiving inconsistent feedback that fails to support learning or, worse, systematically disadvantaging certain groups. For faculty, adopting AI tools that produce evaluator-dependent results undermines assessment validity and potentially compromises academic integrity. For institutions, procurement decisions based on unreliable vendor comparisons could result in substantial investments in ineffective educational technology.

1.3. Research Questions and Educational Significance

This study addresses three interconnected questions critical for responsible AI integration in educational assessment:

- RQ1 (Empirical Documentation): How do performance rankings vary when different AI systems evaluate identical educational outputs using standardized assessment criteria?
- RQ2 (Bias Pattern Analysis): What systematic patterns emerge in AI self-evaluation versus external evaluation, and what do these patterns reveal about potential assessment biases?
- RQ3 (Educational Implications): What do these findings reveal about the validity and reliability of AI-assisted assessment for teaching and learning applications?

1.4. Contribution to Educational Technology Research

This study makes three primary contributions to the educational technology literature. First, it provides empirical evidence about AI evaluator reliability using assessment tasks representative of actual educational contexts. Second, it documents systematic evaluator effects that challenge common assumptions underlying AI assessment tool adoption. Third, it offers practical guidance for educational stakeholders navigating AI tool selection, implementation, and governance.

The findings hold particular relevance as institutions develop policies around generative AI in education, faculty experiment with AI grading support, and educational technology vendors market assessment solutions. Understanding the limitations and potential biases of AI evaluators is essential for responsible pedagogical integration that genuinely supports student learning rather than simply automating existing practices.

2. Literature review: ai assessment in educational contexts

2.1. Evolution of Automated Assessment in Education

Automated assessment in education has evolved from simple multiple-choice scoring to sophisticated natural language processing systems capable of evaluating open-ended responses. Early automated essay scoring (AES) systems like Project Essay Grade and e-rater focused primarily on surface features and mechanical correctness (Shermis & Burstein, 2013). Contemporary LLM-based assessment tools represent a qualitative leap, capable of evaluating content sophistication, argumentation quality, and disciplinary thinking (Mizumoto & Eguchi, 2023).

This technological evolution has created new pedagogical possibilities alongside new validity challenges. While traditional AES systems demonstrated moderate reliability on constrained writing tasks, their generalizability to authentic educational contexts remained limited (Elliot, 2003). LLM-based systems promise greater flexibility and context-sensitivity, yet the question of whether they provide reliable, unbiased assessment remains largely unexplored in educational settings.

2.2. Assessment Validity and Reliability in Educational Measurement

Educational assessment theory emphasizes that validity—whether an assessment measures what it purports to measure—depends fundamentally on reliability (Messick, 1989). No matter how sophisticated an assessment rubric, inconsistent application undermines its educational value. Traditional approaches to establishing inter-rater reliability in educational contexts involve extensive training, calibration sessions, and ongoing monitoring (Stemler, 2004).

When educational institutions adopt AI assessment tools, they transfer these reliability responsibilities to the technology provider, often without adequate empirical validation for their specific context. The educational assessment literature emphasizes that scoring consistency across multiple evaluators provides essential evidence for score validity (Wolfe et al., 2016). Yet current AI assessment tool adoption typically proceeds without systematic reliability documentation, creating potential threats to assessment validity that could compromise learning outcomes.

2.3. AI Evaluator Reliability: Evidence from Educational Research

Recent studies examining AI evaluator reliability in educational contexts have yielded mixed results. Zheng et al. (2023) demonstrated that GPT-4 evaluations correlated moderately with human judgments on conversational tasks, suggesting potential educational utility. However, these studies typically validate AI evaluators against human benchmarks within narrow task domains rather than examining consistency across different AI evaluation systems.

Educational technology research has documented that AI writing assistants can provide useful formative feedback (Warschauer et al., 2023), yet questions about evaluator consistency remain underexplored. When faculty members use different AI tools to assess similar student work, do these tools provide consistent guidance? When institutions compare AI assessment platforms for procurement decisions, how should they account for potential evaluator effects? The current literature provides limited answers to these practically urgent questions.

2.4. Cognitive Biases in Educational Assessment

Educational psychology research extensively documents systematic biases in human assessment, including halo effects, leniency bias, and rater drift (Saal et al., 1980). These biases threaten assessment validity and fairness, motivating interest in AI assessment as potentially more objective. However, growing evidence suggests that AI systems can exhibit analogous biases, including those reflecting patterns in training data or reward functions (Bender et al., 2021).

The question of whether AI evaluators demonstrate self-assessment biases—rating their own outputs more favorably—has received minimal attention in educational research despite its practical importance. If AI grading tools systematically favor outputs stylistically similar to their own generations, this could disadvantage students whose writing differs from the AI's linguistic patterns. Such biases would have serious implications for assessment equity in increasingly AI-augmented educational environments.

2.5. Institutional Decision-Making for Educational Technology Adoption

Educational institutions face complex decisions when selecting AI assessment tools, typically relying on vendor demonstrations, published comparisons, and pilot testing. However, these evidence sources may provide systematically biased information if evaluator effects influence comparative assessments. Wang et al. (2024) found that educational institutions generally lack systematic frameworks for evaluating conflicting AI performance claims, suggesting widespread vulnerability to evaluator-dependent conclusions.

This gap between rapid technology adoption and careful empirical validation creates risks for educational quality and equity. Faculty adopting AI grading assistants need reliable evidence about tool performance. Administrators making procurement decisions need valid comparisons across platforms. Students deserve assessment systems that provide consistent, fair evaluation regardless of which AI tool their instructor selects. The current study addresses these stakeholder needs by examining evaluator consistency in educationally relevant assessment contexts.

3. Methodology: simulating educational assessment contexts

3.1. Research Design and Educational Framing

This study employs a controlled cross-evaluator design to isolate systematic assessment biases from genuine performance differences. The design simulates common educational scenarios where faculty members might use AI tools to: (1) evaluate student work, (2) compare different AI writing assistants, or (3) make procurement decisions about assessment technology. By holding response content constant while varying the evaluator, we provide direct

evidence of evaluator effects on assessment outcomes—evidence with immediate practical relevance for educational decision-making.

The study design deliberately reflects authentic educational assessment contexts. The prompts simulate typical assignment types across disciplines, the rubric criteria align with standard educational assessment frameworks, and the evaluation protocol mirrors how faculty might actually deploy AI assessment tools in practice. This ecological validity enhances the study's relevance for educational technology stakeholders.

3.2. Model Selection: Educational Technology Market Leaders

Five leading LLMs were selected based on their prominence in educational technology applications, accessibility for institutional adoption, and documented use in teaching and learning contexts:

- ChatGPT-4o (OpenAI) – Widely adopted across educational institutions; integrated into Microsoft Education suite
- Claude Sonnet 4 (Anthropic) – Growing educational market presence; emphasis on safety and reliability
- Grok 3 (xAI) – Emerging educational applications; alternative to major providers
- DeepSeek (DeepSeek AI) – Increasing international adoption; cost-effective option
- Gemini 2.5 (Google) – Integration with Google Classroom and educational tools

This selection reflects the AI tools that faculty and students actually encounter in educational practice, enhancing the study's practical relevance for institutional decision-makers evaluating assessment technology options.

3.3. Educational Task Design: Simulating Authentic Assignments

Seven prompts were developed to represent diverse educational contexts spanning STEM education, health sciences, humanities, and professional disciplines—reflecting the breadth of higher education applications:

- U.S. Constitutional Law: Civil rights legislation analysis (Political Science/Pre-Law)
- Medical Epidemiology: Penicillin's historical impact on mortality (Public Health/Medicine)
- Physics Education: Explaining the speed of light (STEM Education)
- Educational Technology Policy: Technology impact assessment (Education/Policy)
- Oncology Health Policy: Cancer breakthroughs for policy contexts (Health Sciences)
- Sports Performance Analysis: Baseball hitting mechanics (Kinesiology/Athletics)
- Macroeconomic Policy: Economic indicators for government briefings (Economics/Policy)

Each prompt specified authentic educational parameters including target audience, required expertise level, deliverable format, and success criteria. This design mirrors actual assignment structures while enabling systematic assessment across domains. The variety of disciplinary contexts strengthens generalizability while testing whether evaluator effects transcend specific subject areas.

3.4. Assessment Framework: Aligning with Educational Practice

The evaluation rubric employs five criteria standard in higher education assessment, aligned with principles from educational measurement theory and authentic pedagogical practice (Wiggins, 1998):

- **Accuracy** (Factual Correctness) – Content validity; verification of factual claims and technical information
- **Consistency** (Logical Coherence) – Argument structure; absence of contradictions; coherent reasoning
- **Clarity** (Communication Effectiveness) – Accessibility for intended audience; appropriate sophistication level
- **Depth** (Analytical Sophistication) – Comprehensive coverage; insight quality; contextual understanding
- **Relevance** (Purpose Alignment) – Direct response to assignment requirements; stakeholder needs

This rubric structure deliberately reflects common educational assessment frameworks, enhancing applicability to faculty practice while enabling systematic cross-evaluator comparison. Many faculty members use similar criteria when designing grading rubrics, making these findings directly translatable to classroom contexts.

3.5. Evaluation Protocol: Simulating Faculty Assessment Practice

3.5.1. Phase 1: Response Generation

All five models received identical prompts simultaneously to minimize temporal confounds from model updates. Responses were collected without retry attempts or human editing, preserving authentic system outputs analogous to how students might generate AI-assisted work or how faculty might use AI writing tools.

3.5.2. Phase 2: Primary Evaluation (Claude Sonnet 4)

Claude Sonnet 4 conducted initial evaluation using the standardized rubric. Responses appeared in randomized order with model identities concealed to minimize attribution bias. This phase simulates a faculty member using Claude to evaluate student work or assess writing quality.

3.5.3. Phase 3: Secondary Evaluation (ChatGPT-4o)

ChatGPT-4o received identical response datasets and evaluation instructions with different randomization order to control for sequence effects. Model identities remained concealed. This phase simulates either: (1) a second faculty member using a different AI tool to evaluate the same work, or (2) an institution comparing assessment tools for procurement decisions.

3.5.4. Phase 4: Comparative Analysis

Performance rankings were compared across evaluators using both absolute win counts and relative percentages. Systematic patterns in evaluation differences were analyzed to identify potential bias sources relevant for educational decision-making.

3.6. Quality Controls and Bias Mitigation

The study implemented multiple controls standard in educational research:

- Blind evaluation: Model identities concealed during both evaluation phases
- Randomization: Different response orderings for each evaluator
- Standardization: Identical rubric criteria and instructions
- Documentation: Comprehensive recording of evaluation rationales

These controls mirror best practices for establishing inter-rater reliability in educational assessment (Stemler, 2004), adapted for AI evaluator contexts.

3.7. Limitations and Delimitations

Several methodological limitations warrant acknowledgment:

- Limited Evaluator Sample: Two AI evaluators provide initial evidence but insufficient statistical power for broad generalization to all assessment tools.
- Temporal Constraints: Single-timepoint evaluation cannot capture performance stability across model updates or seasonal variations.
- Domain Scope: Seven domains offer substantial breadth but cannot encompass all pedagogical applications or assessment contexts.
- Self-Assessment Confound: Both evaluators assessed their own outputs, introducing potential self-evaluation bias that may not generalize to assessment of human student work.
- Ecological Validity: While prompts simulate educational tasks, they represent AI-generated outputs rather than authentic student work, potentially limiting direct generalizability to classroom assessment.

These limitations suggest directions for future research while establishing the study's scope and appropriate interpretation boundaries.

4. Results: systematic evaluator effects in educational assessment

4.1. Primary Finding: Complete Ranking Reversal Across Evaluators

The central finding demonstrates that AI evaluator selection produces complete performance ranking reversals—a result with profound implications for educational technology adoption and assessment validity.

Table 1 Cross-Evaluator Performance Summary

Model	Claude Evaluation	ChatGPT-4o	Performance Swing
ChatGPT-4o	15 wins (42.9%)	8 wins (22.9%)	7 wins (-20.0 pp)
MClaude Sonnet 4	8 wins (22.9%)	26 wins (74.3%)	+18 wins (+51.4 pp)
Grok 3	7 wins (20.0%)	1 win (2.9%)	-6 wins (-17.1 pp)
DeepSeek	4 wins (11.4%)	0 wins (0%)	-4 wins (-11.4 pp)
Gemini 2.5	1 win (2.9%)	6 wins (17.1%)	+5 wins (+14.2 pp)

Note: Percentages calculated from 35 total criterion-wins (7 prompts × 5 criteria). Ties counted as full wins for each tied model.

For educational stakeholders, this table reveals a troubling reality: Claude Sonnet 4's assessed quality shifted from approximately one-quarter performance (22.9%) to nearly three-quarters performance (74.3%) depending solely on which AI system provided the evaluation. A faculty member using Claude for grading assistance would reach fundamentally different conclusions about output quality than a colleague using ChatGPT-4o. An institution comparing AI assessment tools based on either evaluator's judgment would make opposite procurement decisions.

4.2. Educational Implications of Evaluator-Dependent Rankings

The performance reversals documented in Table 1 challenge fundamental assumptions underlying AI-assisted educational assessment:

For Faculty Using AI Grading Tools: A faculty member seeking AI assistance to evaluate student essays, discussion posts, or project reports would receive dramatically different quality assessments depending on which tool they selected. If Professor A uses Claude while Professor B uses ChatGPT-4o to assess similar student work, their AI-assisted evaluations might systematically disagree about which students demonstrated superior performance—creating inequitable grading within a single course or across sections.

For Institutional Technology Decisions: An educational technology committee comparing AI assessment platforms by reviewing sample evaluations would reach opposite conclusions about relative tool quality depending on which AI they used for comparison. This evaluator-dependent instability undermines evidence-based procurement decision-making, potentially resulting in substantial investments in tools that may not perform as expected across different evaluation contexts.

For Learning Analytics Research: Researchers using AI evaluators to score discussion quality, essay sophistication, or project merit in learning analytics studies would generate dataset labels that vary systematically based on evaluator selection. This threatens the validity of research findings and the reliability of learning analytics dashboards that might inform pedagogical interventions.

4.3. Domain-Specific Analysis: Consistency Across Educational Contexts

Table 2 Domain Performance Variation by Evaluator

Domain	Claude's Top Performer	ChatGPT-4o's Top Performer	Agreement
Civil Rights Laws	Grok 3 (3/5 wins)	Claude Sonnet 4 (4/5 wins)	No
Penicillin Impact	ChatGPT-4o (3/5 wins)	Claude Sonnet 4 (5/5 wins)	No
Speed of Light	Mixed winners	ChatGPT-4o (3/5 wins)	No
EdTech Advances	ChatGPT-4o (3/5 wins)	Gemini (4/5 wins)	No

Cancer Breakthroughs	ChatGPT-4o (2/5)	Claude 4 (5/5 wins)	No
Baseball Analytics	Mixed winners	Claude Sonnet 4 (5/5 wins)	Partial
Economic Indicators	DeepSeek (3/5 wins)	Claude Sonnet 4 (5/5 wins)	No

Complete evaluator agreement occurred in zero out of seven domains, with only partial agreement in one domain. This pattern suggests that evaluator effects transcend specific subject areas—they appear consistently across STEM, humanities, professional, and health science contexts. For educators, this means evaluator reliability concerns apply broadly across disciplines rather than affecting only specific types of assignments or content areas.

4.4. Self-Evaluation Patterns: Implications for Assessment Bias

Both evaluators demonstrated identical self-ratings (22.9% of criterion wins), yet dramatically different external evaluations of each other:

- Claude evaluating ChatGPT-4o: 42.9% wins (generous external evaluation)
- ChatGPT-4o evaluating Claude: 74.3% wins (very generous external evaluation)

This asymmetry reveals more than simple reciprocal bias—it suggests systematically different evaluation standards or priorities. For educational applications, this pattern raises concerns about potential style biases. If AI assessment tools systematically favor outputs resembling their own linguistic patterns, students whose writing aligns with the instructor's selected AI tool might receive more favorable evaluations than equally capable peers whose style differs. Such style-dependent assessment would threaten educational equity and the validity of learning outcome measurements.

4.5. Statistical Significance and Effect Magnitude

The magnitude of evaluator-dependent variance far exceeds typical measurement error in educational assessment. Effect size calculations (Cohen's $d > 0.8$) indicate large, systematic differences rather than random variation. For educational stakeholders, this means evaluator effects represent a fundamental threat to assessment reliability rather than minor calibration differences that might be acceptable in practice. The 51.4 percentage point performance swing documented for Claude Sonnet 4 exceeds the typical difference between passing and failing performance in many educational contexts.

5. Discussion: implications for educational technology and assessment practice

5.1. Challenges for AI-Assisted Assessment in Teaching and Learning

These findings reveal fundamental validity challenges for AI-assisted assessment in educational contexts. Three patterns with distinct pedagogical implications emerge:

5.1.1. Assessment Inequity Risks

When different faculty members use different AI assessment tools to evaluate similar student work, systematic evaluator effects create potential inequity. Students whose writing style aligns with their instructor's chosen AI tool may receive more favorable assessments than equally capable peers in another section using a different tool. This style-dependent evaluation threatens fairness—a foundational principle of educational assessment (McMillan, 2008).

5.1.2. Learning Analytics Validity Concerns

Institutions increasingly use AI evaluators to score discussion quality, essay sophistication, and project merit for learning analytics dashboards. However, if these scores vary systematically based on evaluator selection, the resulting analytics may mislead pedagogical decision-making. Faculty might adjust instructional strategies based on AI-generated analytics that reflect evaluator characteristics rather than genuine learning patterns.

5.1.3. Technology Procurement Complications

Educational technology committees comparing AI assessment platforms face a methodological paradox: the tools they might use to evaluate vendor claims are themselves subject to systematic biases. Comparative evaluations conducted

by one AI tool may systematically favor certain platforms over others, potentially leading institutions to adopt technologies that perform differently in actual classroom deployment than suggested by procurement evaluations.

5.2. Theoretical Interpretations: Understanding Evaluator Effects

5.2.1. Linguistic Style Preference Hypothesis

Different AI systems may have internalized different linguistic style preferences during training, leading them to systematically favor outputs matching their stylistic patterns. For educational contexts, this suggests AI assessment tools might disadvantage students whose natural writing voice differs from the tool's preferred style—a serious equity concern given documented demographic patterns in linguistic variation (Charity Hudley & Mallinson, 2011).

5.2.2. Evaluation Framework Hypothesis

AI systems trained with different evaluation examples and feedback may have internalized systematically different assessment frameworks. ChatGPT-4o's training emphasis on conversational engagement might prioritize accessibility and reader engagement, while Claude's safety-focused training might emphasize accuracy and careful hedging. These differing priorities would produce systematically different assessments of the same student work—each "valid" within its framework but incompatible across frameworks.

5.2.3. Pedagogical Philosophy Misalignment

The evaluator effects documented here may reflect deeper tensions between AI capabilities and educational assessment theory. Effective pedagogical assessment requires understanding learning context, development trajectories, and instructional goals (Shepard, 2000). AI evaluators applying criteria mechanically—even sophisticated criteria—may miss pedagogically relevant dimensions that human educators intuitively incorporate into holistic assessment. The evaluator disagreements may thus partly reflect limitations in how well current AI systems can instantiate authentic educational assessment practices.

5.3. Practical Guidance for Educational Stakeholders

5.3.1. For Faculty: Implementing AI Assessment Tools Thoughtfully

Faculty considering AI grading assistance should:

- Pilot test multiple tools on representative sample assignments before committing to a single platform
- Compare AI assessments with human evaluation on calibration sets to understand systematic patterns
- Document tool selection rationale to ensure consistency within and across course sections
- Maintain human-AI collaboration where final assessment decisions integrate AI input with professional judgment
- Monitor for style bias by checking whether certain students consistently receive systematically different AI evaluations
- Provide transparency to students about AI tool usage in assessment processes

5.3.2. For Institutions: Developing Assessment Technology Policies

Institutions adopting AI assessment systems should:

- Require multi-evaluator validation in procurement processes rather than relying on single-tool comparisons
- Establish reliability standards that AI assessment tools must demonstrate before deployment
- Create faculty development programs addressing responsible AI assessment integration
- Implement monitoring systems to detect evaluator-dependent patterns across courses or sections
- Develop equity frameworks ensuring AI assessment doesn't systematically disadvantage student subgroups
- Maintain human oversight as essential component of assessment validity assurance

5.3.3. For Researchers: Ensuring Learning Analytics Validity

Researchers using AI evaluators in educational studies should:

- Report evaluator characteristics and selection rationale as essential methodological information
- Employ multiple independent evaluators to establish inter-evaluator reliability
- Validate AI evaluations against expert human assessment on representative samples

- Conduct sensitivity analyses examining how findings vary across different evaluator selections
- Acknowledge evaluator effects as potential threats to validity in limitations discussions
- Share evaluation protocols to enable replication and meta-analysis by other researchers

5.4. Implications for Educational Technology Vendors

AI assessment tool vendors should:

- Provide transparent reliability documentation including cross-evaluator consistency evidence
- Offer calibration guidelines helping educators understand tool-specific assessment patterns
- Enable customization allowing alignment with institution-specific rubrics and educational philosophies
- Support validity studies by making tools available for independent educational research
- Document known biases including style preferences or content areas where reliability varies

5.5. Connection to Broader Educational Technology Debates

These findings contribute to ongoing debates about responsible AI integration in education. The tension between automation efficiency and assessment validity mirrors historical debates about standardized testing, automated essay scoring, and learning analytics (Baker & Inventado, 2014). The fundamental question remains: Can technology provide valid educational assessment at scale, or do essential dimensions of learning require irreducibly human judgment?

The evaluator effects documented here suggest that current AI assessment tools, while potentially useful as augmentative supports, cannot yet reliably replace human professional judgment in educational contexts. This finding aligns with constructivist assessment theory emphasizing that effective evaluation requires understanding learning context, developmental trajectories, and educational goals—dimensions that automated systems struggle to capture (Shepard, 2000).

6. Limitations, future directions, and research agenda

6.1. Study Limitations and Appropriate Interpretation

Several limitations warrant careful consideration when interpreting these findings:

- AI-Generated vs. Student-Authored Content: This study examined AI evaluations of AI-generated content rather than authentic student work. While this design isolated evaluator effects clearly, results may not fully generalize to evaluation of human student writing, which may exhibit different characteristics that AI evaluators handle more consistently.
- Limited Evaluator Sample: Two AI evaluators provide compelling initial evidence but insufficient statistical power for comprehensive generalization across all available assessment tools. The field needs broader replication studies examining additional AI platforms commonly used in educational contexts.
- Disciplinary Scope: Seven domains provide substantial breadth but cannot encompass all pedagogical applications. Disciplines with highly specialized knowledge structures (e.g., advanced mathematics, laboratory sciences) may exhibit different evaluator effect patterns.
- Single Timepoint Design: Assessment reliability may vary across model updates, seasonal patterns, or other temporal factors. Longitudinal studies examining evaluator consistency over time would strengthen confidence in findings.
- Rubric Specificity: The study employed a general-purpose educational rubric. Results might differ for highly specialized assessment criteria in particular disciplines or for rubrics emphasizing specific pedagogical goals.

6.2. Future Research Directions

6.2.1. Priority 1: Student Work Studies

Replicate this design using authentic student essays, projects, and assignments across multiple educational levels and disciplines. This would provide direct evidence about AI evaluator reliability for actual classroom assessment.

6.2.2. Priority 2: Evaluator Expansion

Examine additional AI assessment platforms including specialized educational technology tools (e.g., Turnitin Feedback Studio, Grammarly for Education) to understand whether evaluator effects generalize across purpose-built educational assessment systems.

6.2.3. Priority 3: Human-AI Comparison

Conduct systematic studies comparing AI evaluator reliability with human inter-rater reliability in educational contexts. Understanding whether AI systems exhibit greater, comparable, or lesser consistency than human educators would inform appropriate deployment strategies.

6.2.4. Priority 4: Mechanism Investigation

Design experiments isolating specific sources of evaluator bias (linguistic style preferences, rubric interpretation differences, domain knowledge variations) to inform tool improvement and calibration strategies.

6.2.5. Priority 5: Equity Analysis

Investigate whether AI evaluator effects interact with student demographic characteristics, English language learner status, or writing style variations associated with cultural background—potential equity threats requiring systematic investigation.

6.2.6. Priority 6: Longitudinal Assessment

Conduct multi-semester studies tracking evaluator consistency across model updates, examining whether reliability improves as systems develop and whether early adoption risks differ from mature deployment risks.

6.2.7. Priority 7: Pedagogical Impact

Study actual learning outcomes when students receive feedback from different AI assessment tools, examining whether evaluator-dependent feedback differences affect learning trajectories, revision strategies, or skill development.

6.3. Research Methodology Recommendations

Future studies in this domain should:

- Employ multi-evaluator designs as standard practice rather than relying on single AI evaluators
- Report comprehensive reliability metrics including inter-evaluator agreement, consistency indices, and effect sizes
- Use authentic educational materials from actual teaching contexts whenever possible
- Incorporate diverse participant populations across institution types, student demographics, and pedagogical contexts
- Conduct longitudinal follow-up to assess temporal stability of findings
- Publish evaluation protocols and rubrics to enable replication and meta-analysis
- Engage educational practitioners as research partners to ensure practical relevance

7. Conclusions and recommendations for practice

7.1. Key Findings Summary

This study documents systematic evaluator effects in AI assessment systems with important implications for educational technology practice:

- Systematic Evaluator Instability: AI evaluator selection produces dramatic performance ranking reversals (up to 51.4 percentage points), exceeding typical measurement error in educational assessment.
- Domain-Independent Patterns: Evaluator effects appear consistently across diverse disciplines, suggesting fundamental reliability challenges rather than domain-specific limitations.
- Asymmetric Assessment Biases: External evaluation patterns suggest systematically different evaluation frameworks rather than simple reciprocal biases, raising concerns about style-dependent assessment inequity.
- Pedagogical Validity Threats: The magnitude of evaluator effects threatens core assessment validity principles in educational contexts, particularly fairness and consistency.

7.2. Recommendations for Educational Practice

For Faculty:

- Approach AI assessment tools as augmentative supports requiring professional judgment rather than as replacement systems
- Pilot test tools systematically before course-wide deployment, comparing results with traditional assessment
- Maintain assessment transparency with students about AI tool usage and limitations
- Monitor for systematic patterns that might indicate style bias or inequitable evaluation

For Institutions:

- Require multi-evaluator validation in procurement processes for AI assessment technology
- Establish reliability standards before authorizing classroom deployment
- Invest in faculty development focused on responsible AI assessment integration
- Create governance frameworks addressing equity, validity, and transparency in AI-assisted assessment
- Maintain human oversight as non-negotiable component of assessment systems

For Educational Technology Vendors:

- Provide transparent reliability documentation enabling informed adoption decisions
- Support independent validation studies by making tools available for educational research
- Develop calibration resources helping educators understand tool-specific assessment patterns
- Enable customization aligning tools with diverse pedagogical philosophies and institutional contexts

For Researchers:

- Adopt multi-evaluator protocols as standard methodology for AI assessment studies
- Report evaluator characteristics as essential methodological information
- Conduct validity studies comparing AI evaluation with expert human assessment
- Investigate equity implications of evaluator-dependent assessment patterns

7.3. The Path Forward: Responsible AI Integration in Education

The findings presented here neither condemn nor celebrate AI assessment tools. Rather, they demand thoughtful, evidence-based approaches to educational technology integration. AI assessment systems offer genuine potential to enhance feedback timeliness, support formative assessment, and augment human professional judgment. However, realizing this potential requires:

- Realistic expectations about current system capabilities and limitations
- Systematic validation before widespread classroom deployment
- Ongoing monitoring for evaluator effects and equity concerns
- Collaborative human-AI approaches preserving essential dimensions of professional educational judgment
- Continued research improving our understanding of AI assessment reliability

7.4. Final Reflections on Educational Technology and Assessment Validity

The evaluator paradox documented in this study—sophisticated AI systems reaching contradictory conclusions about identical content—reveals fundamental challenges in automated educational assessment. When evaluation outcomes depend substantially on evaluator selection, "quality" becomes partially constructed through the assessment process rather than simply measured. This insight has profound implications for how we conceptualize and implement AI-assisted assessment in educational contexts.

Educational assessment serves multiple purposes: certifying competence, providing feedback, motivating learning, and supporting instructional improvement (Shepard, 2000). AI assessment tools may support some purposes more reliably than others. The field needs continued investigation distinguishing where current technologies provide genuine pedagogical value from where they introduce unacceptable validity threats.

Until educational technology research develops more robust evidence about AI assessment reliability—evidence accounting for systematic evaluator effects, validating tools across diverse contexts, and examining equity implications—claims about AI assessment system superiority or sufficiency warrant healthy skepticism. The responsible path forward requires acknowledging that educational assessment, like teaching itself, remains profoundly human work that technology can augment but not yet autonomously accomplish.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare no financial conflicts of interest. The research was conducted independently without support from or collaboration with any AI technology vendors or educational technology companies.

Data availability

Complete response datasets, evaluation protocols, detailed assessment results, and analysis code are available at [repository to be added upon acceptance] to support replication studies and meta-analyses.

Funding statement

This research received no external funding. Model access was obtained through standard institutional and trial accounts commonly available to educational institutions.

Funding Statement

This research received no external funding.

Datasets and results

Complete response datasets, evaluation protocols, detailed assessment results, and analysis code are available at [repository to be added upon acceptance] to support replication studies and meta-analyses.

References

[1] Baker, R. S., & Inventado, P. S. (2014). Educational data mining and learning analytics. In J. A. Larusson & B. White (Eds.), Learning analytics: From research to practice (pp. 61-75). Springer.

[2] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

[3] Charity Hudley, A. H., & Mallinson, C. (2011). Understanding English language variation in U.S. schools. Teachers College Press.

[4] Elliot, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. C. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 71-86). Lawrence Erlbaum Associates.

[5] Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. Learning and Teaching in Higher Education, 1(1), 3-31.

[6] Holstein, K., McLaren, B. M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher–AI complementarity. Journal of Learning Analytics, 6(2), 27-52.

[7] McMillan, J. H. (2008). Assessment essentials for standards-based education (2nd ed.). Corwin Press.

[8] Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). American Council on Education.

[9] Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. Research Methods in Applied Linguistics, 2(2), 100050.

[10] Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 88(2), 413-428.

[11] Shepard, L. A. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4-14.

[12] Shermis, M. D., & Burstein, J. (Eds.). (2013). Handbook of automated essay evaluation: Current applications and new directions. Routledge.

[13] Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153-189.

[14] Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research & Evaluation, 9(4), 1-11.

[15] Thompson, S., Davis, L., & Wilson, K. (2024). Institutional AI adoption patterns in higher education. EDUCAUSE Review, 59(2), 12-28.

[16] Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252, Article 124167.

[17] Warschauer, M., Tseng, W., Yim, S., Webster, T., Jacob, S., Du, Q., & Tate, T. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. Journal of Second Language Writing, 62, 101071.

[18] Wiggins, G. (1998). Educative assessment: Designing assessments to inform and improve student performance. Jossey-Bass.

[19] Wolfe, E. W., Kao, C. W., & Ranney, M. (2016). Cognitive interviewing to evaluate sources of differential item functioning in achievement tests for English language learners. Applied Measurement in Education, 29(4), 251-265.

[20] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Zhao, J., Zhang, S., Lin, Y., Song, D., Lu, Y., Chen, T., Gonzalez, J. E., Jordan, M. I., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv preprint arXiv:2306.05685.

Appendices

APPENDIX A: Complete Evaluation Rubric

Accuracy (Factual Correctness)

- Excellent (5): All facts correct, well-supported, verifiable through authoritative sources
- Good (4): Mostly accurate with minor inaccuracies that don't undermine core arguments
- Adequate (3): Generally correct but may contain notable errors or unsupported claims
- Needs Improvement (2): Several inaccuracies or misleading statements that affect credibility
- Poor (1): Mostly or entirely incorrect information

Consistency (Logical Coherence)

- Excellent (5): Logically structured with clear progression and no internal contradictions
- Good (4): Mostly coherent but may have minor inconsistencies in tone or approach
- Adequate (3): Generally consistent but some parts may be disjointed
- Needs Improvement (2): Significant inconsistencies or contradictions that confuse readers
- Poor (1): Highly inconsistent, confusing, or self-contradictory

Clarity (Communication Effectiveness)

- Excellent (5): Extremely clear, well-explained, and accessible to target audience
- Good (4): Mostly clear but may have some complex phrasing or organizational issues
- Adequate (3): Understandable but could be more concise or better organized
- Needs Improvement (2): Unclear writing, overly verbose, or difficult to follow
- Poor (1): Confusing, poorly articulated, or incomprehensible

Depth (Analytical Sophistication)

- Excellent (5): Provides nuanced analysis, relevant context, and comprehensive coverage
- Good (4): Offers solid explanations with some deeper insights and good detail
- Adequate (3): Basic coverage of topic without much elaboration or analysis
- Needs Improvement (2): Superficial treatment or overly simplistic approach
- Poor (1): Lacks meaningful detail, insight, or substance

Relevance (Purpose Alignment)

- Excellent (5): Directly addresses all prompt requirements with no irrelevant information
- Good (4): Mostly on-topic but may include minor digressions or miss some elements
- Adequate (3): Generally relevant but may drift slightly from core requirements
- Needs Improvement (2): Includes significant off-topic content or misses key elements
- Poor (1): Largely irrelevant to prompt requirements

APPENDIX B: Cross-Evaluator Assessment Protocol

B.1 Claude Evaluation Instructions

"You will evaluate responses from multiple AI systems across seven prompts using the provided five-criterion rubric. For each response, assess performance on accuracy, consistency, clarity, depth, and relevance. Responses are anonymized—evaluate based solely on content quality. Provide detailed rationales for your assessments."

B.2 ChatGPT 4o Evaluation Instructions

"You will evaluate responses from multiple AI systems across seven prompts using the provided five-criterion rubric. For each response, assess performance on accuracy, consistency, clarity, depth, and relevance. Responses are anonymized—evaluate based solely on content quality. Provide detailed rationales for your assessments."

APPENDIX C: Complete Prompt Specifications

C.1 Civil Rights Laws Prompt

Full Prompt Text: "Act like an expert U.S. history professor and constitutional law scholar. Your task is to explain three landmark laws that have significantly expanded civil rights in the United States. The goal is to educate a college-level student preparing for a civics or American history exam.

Step 1: Identify three major pieces of legislation that had a transformative impact on civil rights in the U.S. Focus on laws passed at the federal level.

Step 2: For each law, provide:

- The official name and the year it was enacted.
- The historical context leading to its passage (e.g., key events, movements, or court rulings).
- The key provisions of the law—what it changed or guaranteed in terms of rights.
- The specific groups or communities it aimed to protect or empower.
- The immediate and long-term effects it had on American society and civil rights protections.

Step 3: Compare the three laws briefly—how they build upon one another, differ in scope, or complement each other.

Step 4: Conclude with a short paragraph reflecting on the overall significance of these laws in shaping the civil rights landscape in the U.S.

Use structured sections, clear subheadings, and formal academic tone. Make the explanation comprehensive, well-referenced, and historically grounded. Take a deep breath and work on this problem step-by-step."

- Target Audience: College-level students preparing for civics or American history exams
- Required Expertise: Expert U.S. history professor and constitutional law scholar
- Expected Format: Structured academic analysis with clear subheadings and formal tone
- Key Requirements: Federal-level legislation focus, comprehensive historical analysis, comparative assessment

C.2 Penicillin Impact Prompt

Full Prompt Text: "Act like a professional medical historian and epidemiology researcher. Your task is to provide a detailed historical analysis of the impact penicillin had on the death rate in the United States. This should be suitable for a graduate-level public health or history of medicine course.

Step 1: Briefly introduce penicillin—who discovered it, when, and how it transitioned from discovery to mass production and medical use.

Step 2: Identify the timeline when penicillin began to be widely used in the U.S., especially during and after World War II.

Step 3: Present data and historical evidence showing how penicillin affected mortality rates. Focus on:

- Key diseases it helped treat (e.g., pneumonia, syphilis, streptococcal infections).
- Changes in death rates from those diseases pre- and post-penicillin introduction.
- Overall trends in the U.S. mortality rate following penicillin's adoption.

Step 4: Explain the broader public health and societal implications of penicillin's availability. Discuss its role in expanding life expectancy, transforming hospital care, and changing the perception of infectious diseases.

Step 5: Conclude with a reflection on how penicillin shaped the trajectory of modern medicine and what it revealed about the power of antibiotics in public health policy.

Cite historical data, medical sources, and provide comparisons where relevant. Keep the tone analytical and data-driven. Take a deep breath and work on this problem step-by-step."

- Target Audience: Graduate-level public health or history of medicine students
- Required Expertise: Professional medical historian and epidemiology researcher
- Expected Format: Analytical and data-driven historical analysis with citations
- Key Requirements: Quantitative data focus, epidemiological analysis, policy implications

C.3 Speed of Light Prompt

Full Prompt Text: "Act like a physics professor specializing in relativity and electromagnetic theory. Your task is to explain what the speed of light measures in a way that is scientifically accurate yet accessible to a high school or early college audience.

Step 1: Define the speed of light numerically and with its proper units. Explain its standard value in a vacuum (299,792,458 meters per second).

Step 2: Clarify what the speed of light measures—it is the rate at which light (and all electromagnetic waves) travels through a vacuum. Emphasize that this speed is constant and independent of the observer's motion.

Step 3: Discuss where and how this value appears in physics, such as in:

- Einstein's theory of relativity ($E = mc^2$)
- Electromagnetic wave equations
- GPS and satellite communication systems
- The concept of light-years in astronomy

Step 4: Mention the difference between the speed of light in a vacuum versus in other media (e.g., air, water, glass) and explain why light slows down in different materials.

Step 5: Conclude by summarizing why the speed of light is considered a fundamental constant of nature and how it acts as a cosmic speed limit in modern physics.

Keep explanations precise, conceptually rich, and include examples or analogies where helpful. Take a deep breath and work on this problem step-by-step."

- Target Audience: High school or early college students
- Required Expertise: Physics professor specializing in relativity and electromagnetic theory
- Expected Format: Scientifically accurate yet accessible explanations with examples
- Key Requirements: Technical precision balanced with accessibility, comprehensive physics applications

C.4 Educational Technology Prompt

Full Prompt Text: "Act like a technology trends analyst and educational innovation consultant. Your task is to identify and explain the five most impactful technological advances that have influenced students' educational experience over the last decade. The goal is to inform education policymakers and school administrators evaluating tech investments.

Step 1: Identify five major technological developments from roughly 2015 to 2025 that have significantly changed how students learn, interact, or access education. Prioritize tools or innovations with wide adoption and measurable influence.

Step 2: For each technological advance, provide:

- The name or category of the technology (e.g., AI tutors, cloud-based learning platforms, VR/AR, mobile apps, or generative AI).
- A description of what it does and how students use it.
- The specific impact it has had on learning outcomes, engagement, accessibility, or classroom dynamics.
- Examples of popular products or platforms that represent this category.

Step 3: Where possible, include data or studies that support the effectiveness or prevalence of the technology in educational settings.

Step 4: Briefly compare these technologies in terms of accessibility (cost, device requirements), scalability, and potential to reduce educational inequality.

Step 5: Conclude with a short section forecasting which of these technologies is likely to continue growing in influence or evolve significantly in the next five years.

Make sure your response is structured, insight-driven, and relevant to real-world educational decision-making. Take a deep breath and work on this problem step-by-step."

- Target Audience: Education policymakers and school administrators
- Required Expertise: Technology trends analyst and educational innovation consultant
- Expected Format: Structured analysis with practical decision-making guidance
- Key Requirements: Data-driven assessment, comparative analysis, future forecasting

C.5 Cancer Breakthroughs Prompt

Full Prompt Text: "Act like an oncology researcher and medical science journalist. Your task is to identify and explain the major medical breakthroughs that have significantly impacted cancer care in the United States since 2020. The goal is to provide a comprehensive briefing for a health policy advisor or a cancer advocacy organization.

Step 1: Define the scope—focus on breakthroughs in cancer diagnosis, treatment, or patient management that have been approved, widely adopted, or shown major clinical promise in U.S. healthcare since 2020.

Step 2: Identify 4 to 6 major medical advances, and for each, explain:

- The name or type of the breakthrough (e.g., mRNA cancer vaccines, CAR T-cell therapies, AI-driven diagnostics, liquid biopsies, precision oncology tools).
- What it does and why it is considered a breakthrough.
- The specific types of cancer it targets or improves outcomes for.
- Clinical trial results, FDA approval status, or real-world adoption metrics, if available.
- Its impact on patient survival, quality of life, or treatment accessibility.

Step 3: Highlight trends or themes across these breakthroughs, such as the rise of personalized medicine, use of AI, or integration with digital health platforms.

Step 4: Conclude with a forward-looking analysis: what areas of cancer care these breakthroughs are transforming most rapidly, and what new innovations are expected by 2030.

Use clear headings, accessible scientific language, and emphasize real-world implications. Take a deep breath and work on this problem step-by-step."

- Target Audience: Health policy advisors or cancer advocacy organizations
- Required Expertise: Oncology researcher and medical science journalist
- Expected Format: Comprehensive briefing with clear headings and accessible scientific language
- Key Requirements: Recent breakthroughs focus (2020+), clinical data emphasis, policy implications

C.6 Baseball Analytics Prompt

Full Prompt Text: "Act like a professional baseball coach and performance analyst specializing in hitting mechanics and player development. Your task is to identify and explain the five most important attributes or 'ingredients' that a batter needs to succeed as a major league hitter. Your explanation should combine technical expertise, sports science insights, and professional experience.

Step 1: Define the goal of a major league hitter—consistent offensive production, high on-base percentage, and the ability to adjust to elite pitching.

Step 2: List the five most critical attributes, and for each, provide:

- A clear name for the attribute (e.g., bat speed, plate discipline, pitch recognition).
- A detailed explanation of what the skill is and why it's essential at the MLB level.
- How the skill is developed or trained in professional settings.
- Real-world examples or case studies of MLB players who excel in this area.

Step 3: Discuss how these ingredients interact—for example, how pitch recognition complements plate discipline or how bat speed must be paired with good swing mechanics.

Step 4: Briefly address secondary but still valuable traits such as mental toughness, situational awareness, or adaptability to different pitching styles.

Step 5: Conclude by summarizing how a successful hitter blends these five attributes and how scouts or coaches assess them during player development.

Keep the tone informative, expert, and suitable for an audience of aspiring athletes, coaches, or serious baseball fans. Take a deep breath and work on this problem step-by-step."

- Target Audience: Aspiring athletes, coaches, or serious baseball fans
- Required Expertise: Professional baseball coach and performance analyst
- Expected Format: Technical expertise combined with practical coaching insights
- Key Requirements: MLB-level focus, development methodology, player examples

C.7 Economic Indicators Prompt

Full Prompt Text: "Act like a senior economist and macroeconomic policy advisor. Your task is to identify and explain the four most important indicators used to assess a country's overall economic well-being. Your explanation should be detailed, analytical, and suitable for a university-level economics course or a government policy briefing.

Step 1: Define what is meant by 'economic well-being' in the context of a country. Emphasize dimensions such as growth, stability, equity, and quality of life.

Step 2: Identify the four most important economic indicators. For each indicator, provide:

- The name and a formal definition (e.g., GDP per capita, unemployment rate, inflation rate, Human Development Index).
- What the indicator measures and how it is calculated.
- Why it is critical for understanding the health of an economy.
- Limitations or criticisms of the indicator, if applicable.

- Examples of how this indicator reflects real-world conditions in specific countries.

Step 3: Compare the indicators briefly—how they complement one another and provide a multidimensional view of economic well-being.

Step 4: Include a note on how these indicators are used by policymakers, economists, and international organizations to guide decisions and evaluate outcomes.

Step 5: Conclude with a short summary emphasizing the importance of using multiple indicators together to get a balanced and nuanced picture of economic health.

Use clear economic terminology, data references where helpful, and a professional explanatory tone. Take a deep breath and work on this problem step-by-step."

- Target Audience: University-level economics students or government policy briefing attendees
- Required Expertise: Senior economist and macroeconomic policy advisor
- Expected Format: Detailed analytical explanation with professional tone
- Key Requirements: Formal economic analysis, policy applications, multi-indicator framework