mentafy

Beyond Detection

Process Transparency as the Key to Academic Integrity in the AI Era

Address :

Gottfried-Hagen Straße 60-62
51105 Cologne
Germany

Website :

www.mentafy.com
info@mentafy.com

Phone :

+49 (221) 56070319

Table Of Contents

Beyond Detection: Process Transparency as the Key to Academic Integrity in the AI Era

Executive Summary

Academic integrity faces a pivotal challenge in the era of generative Artificial Intelligence (AI). Tools like ChatGPT can produce human-like essays, rendering traditional plagiarism and 'first generation' AI-detection methods increasingly ineffective and even counter productive. This whitepaper argues that **process transparency** – requiring students to show how their work is produced – is a more sustainable and fair solution.
Key points include:

Limitations of Detection

AI-generated content is difficult to reliably detect for instructors [1]. Studies find current AI-detection tools to be *"neither accurate nor reliable"* [2], with roughly only a 50% success rate on average. False positives repeatedly lead to serious injustices [3].

Process Transparency as Solution

Requiring students to document research, drafts, and any AI assistance (a "proof-of-work" approach) provides evidence of learning and originality [4]. This is aligning with calls to focus on whether learning has occurred rather than catching cheaters [5].

Responsible AI Citation

Transparency requires openly acknowledging AI assistance, like traditional source citations, to maintain academic integrity [6]. Tools like the Artificial Intelligence Assessment Scale (AIAS) [7] provide clear guidelines - from "no AI usage" to "full AI collaboration" - enabling precise alignment of AI involvement and disclosure with learning outcomes.

Mentafy – A Balanced Approach

Mentafy is highlighted as an example solution. It documents the whole work process to give educators evidence of how an assignment was crafted [8]. AI detector flaws are addressed: it highly respects privacy, avoids "black box" judgments, and provides real-time feedback to students.

Recommendations on how to realize academic integrity in the AI era are provided. Updating integrity policies, training stakeholders on ethical AI use, implementing process-oriented documentation tools, and fostering a culture where transparency and authentic learning, rather than detection alone, define integrity [9].

"Anyone, who doesn't take truth seriously in small matters canot be trusted with large ones either", ***Albert Einstein***

Introduction

The rise of generative AI has disrupted traditional notions of academic integrity. Since ChatGPT emerged in late 2022, educators have grappled with preventing AI-assisted cheating, initially favoring bans or advanced detection methods. Yet, such reactionary approaches are increasingly inadequate. A fundamental question arises: Does our definition of academic integrity still apply in the AI era [9]?

Evidence indicates that despite AI's novelty, cheating motivations and rates remain consistent, rooted in familiar academic pressures rather than technology itself [10] [11]. However, educators now struggle more than ever to distinguish genuine student work from AI-generated content, eroding trust essential to educational relationships. While AI can generate competent first drafts, writing skills remain crucial – shifting focus to critical thinking, analysis, and revision.

In this evolving landscape, academic integrity must be redefined. Rather than merely asking, "Was external help used?" educators mast ask, "How was this work produced, and did genuine learning occur?" We will examine why traditional methods fail and explore how process-oriented frameworks like the AI Assessment Scale and tools such as Mentafy can maintain honesty and learning without inhibiting innovation. The goal is to firstly support students in their academic journey and secondly, equip policymakers, educators, and integrity officers with practical strategies to uphold trust and fairness in education.

"94% of AI-cheating is undetected today", *Scarfe et al. 2024*

Why Straight-Forward AI-Detection Fails

For decades, upholding academic integrity in terms of a technological solution often meant catching cheaters through plagiarism checkers and exam proctors. Now in the AI era, many turned to automated **AI detectors** to identify machine-generated text in student work. However, this approach faces severe and compounding limitations:

> ***"AI detectors tested have a mean accuracy rating of only 39.5%",***
> ***Perkins et al., 2024***

Unreliable Accuracy

An international study in 2023 showed that a dozen popular AI detectors were *"neither accurate nor reliable"* in real classroom scenarios [2]. In practice, these tools often miss a massive portion of AI-written text or falsely flag human-written prose [12] [13].

False Positives and Erosion of Trust

AI detectors risk falsely flagging authentic student work as AI-generated, harming students' academic reputations and eroding trust. A high-profile example in Texas [4] showed a professor wrongly failing half his class by using ChatGPT itself as a detector, unaware the tool inaccurately claimed authorship of any text presented to it.

Easy to Evade

Even when detectors do work, students quickly find simple ways to evade them. For instance, paraphrasing AI-generated text can defeat many detection algorithms [2]. Some tools flagged chunks of text that were verbatim from AI, but once the wording was tweaked by humans or even another AI (a process often called *"AI laundering"*), the detectors were largely fooled. As two researchers summing up their tests of various detectors warned: *"we should assume students will be able to break any AI-detection tools, regardless of their sophistication."* [14].

Lack of Transparency ("Black Box" Problem):

Most AI detection algorithms operate opaquely – they give a verdict (or a score) without clear justification. This "black box" nature means neither students nor educators get insight into why a piece of work was flagged. A student said that their essay is 80% likely AI-written is often left bewildered, unable to understand the result. Likewise, instructors cannot in good conscience explain the decision or verify its correctness. This opacity undermines the very purpose of an integrity system.

The Value of Process Transparency

If we cannot reliably detect AI involvement after the fact, the solution is to discourage inappropriate use *before and during* the work process. Process transparency means students openly document how their work was created: including planning, drafts, research notes, and any tools (AI or otherwise) used along the way. This approach shifts the focus from policing the final document to understanding and guiding the steps that led there.

From Policing to Learning

Educators and integrity officers are increasingly advocating moving *"from a focus on detecting cheating to focusing on detecting whether learning has occurred"* [5]. By examining a student's process an instructor can see evidence of genuine learning: the evolution of ideas, the mistakes corrected, and the personal insights added. This addresses the heart of academic integrity: **ensuring students engage with and understand their work.** After all, a perfectly polished essay means little if the student did not learn from creating it.

Proof-of-Work as Deterrent

Requiring a "proof-of-work" – a record of each research and writing stage - deters dishonesty by making cheating harder and riskier. Students must show their process, so copying from AI or others means faking outlines, notes, and drafts, increasing effort and chances of being caught. Sincere students, meanwhile, can clearly show their work. Tools like PowerNotes help document and review this process, enabling feedback and verifying originality. Proof-of-work does not just detect misconduct – it guides students through proper academic practice step-by-step, reducing reliance on shortcuts like AI.

Skill Development and Critical Thinking

As generative AI handles routine drafting, students should focus on what AI cannot: critical analysis, ethical judgment, and creative input. Writing education experts note that *"GenAI's ability to create drafts does not render composition obsolete; it shifts emphasis to critical analysis, revision and 'prompt engineering.'"* [9]. Students showing their edits builds meta-cognitive skills - turning AI into a learning aid under guided, transparent conditions.

Fostering a Culture of Honesty

Process transparency promotes trust and integrity. When instructors clearly address AI use, students are more likely to be honest about their methods, instead of trying to game an algorithm. As with plagiarism, normalizing disclosure reduces the temptation for misuse. For homework institutions can encourage brief AI usage statements (e.g., *"I brainstormed with GPT-4, wrote the text myself, and used Grammarly to proofread"*), helping demystify AI.

Introducing the Artificial Intelligence Assessment Scale (AIAS)

One practical framework emerging from the push for transparency and guided AI integration is the **Artificial Intelligence Assessment Scale (AIAS)**. Developed by education researchers in 2024, the AIAS offers institutions a structured way to decide *how much* AI use is appropriate for a given assignment, in alignment with learning objectives [7].

1	NO AI	The assessment is completed entirely without AI assistance in a controlled environment, ensuring that students rely solely on their existing knowledge, understanding, and skills **You must not use AI at any point during the assessment. You must demonstrate your core skills and knowledge.**
2	AI PLANNING	AI may be used for pre-task activities such as brainstorming, outlining and initial research. This level focuses on the effective use of AI for planning, synthesis, and ideation, but assessments should emphasise the ability to develop and refine these ideas independently. **You may use AI for planning, idea development, and research. Your final submission should show how you have developed and refined these ideas.**
3	AI COLLABORATION	AI may be used to help complete the task, including idea generation, drafting, feedback, and refinement. Students should critically evaluate and modify the AI suggested outputs, demonstrating their understanding. **You may use AI to assist with specific tasks such as drafting text, refining and evaluating your work. You must critically evaluate and modify any AI-generated content you use.**
4	FULL AI	AI may be used to complete any elements of the task, with students directing AI to achieve the assessment goals. Assessments at this level may also require engagement with AI to achieve goals and solve problems. **You may use AI extensively throughout your work either as you wish, or as specifically directed in your assessment. Focus on directing AI to achieve your goals while demonstrating your critical thinking.**
5	AI EXPLORATION	AI is used creatively to enhance problem-solving, generate novel insights, or develop innovative solutions to solve problems. Students and educators co-design assessments to explore unique AI applications within the field of study. **You should use AI creatively to solve the task, potentially co-designing new approaches with your instructor.**

Image 1: Overview of the Artificial Intelligence Assessment Scale (AIAS) [7]

By setting these five levels, the AIAS gives educators a flexible yet clear policy toolkit. For each assignment, instructors (or departments) can choose an AIAS level based on what skills or knowledge they are assessing. If the aim is to assess raw writing ability or factual recall, Level 1 is chosen to prohibit AI. If the aim is to assess critical thinking or the ability to improve a draft, a mid-level (2–4) might be selected. This transparency helps students understand expectations and limits.

The Role of Responsible AI Citation

Even with clear policies on AI usage, one element remains essential to process transparency and academic integrity: **honest disclosure and citation of AI assistance**. Just as students are expected to cite books, articles, or websites that informed their work, they should also acknowledge the role of AI tools.

AI as a Source or Tool

The academic community is coalescing around the idea that AI-generated content, of significant AI assistance, must be cited or credited in scholarly work [6]. Of course, citing AI presents new challenges: AI outputs are not fixed (ChatGPT might produce a different answer each time) and they are not traditional sources that one can retrieve later.

Preventing Misinterpretation

Mandatory AI citation ensures evaluators are not misled about student capabilities. Declaring AI assistance allows evaluators to adjust assessments appropriately, focusing on students' analysis and editing. Universities are therefore adding explicit policies: *"Any AI-generated content used in assignments must be properly attributed. Failure to disclose AI assistance may result in plagiarism charges."* [15].

Educating How to Cite

Libraries and writing centers have begun creating educational resources, how to cite AI [16] [17]. For instance, university guides often provide examples: citing ChatGPT in MLA might involve a citation like *"OpenAI. ChatGPT response to a prompt about [topic], March 3, 2025."* in the bibliography, whereas APA might suggest an in-text mention and an acknowledgement in the methodology or author's note. Teaching proper AI citation serves a dual purpose: it reinforces to students that using AI is not "free help" without strings – they must take responsibility for it – and it provides a trail that instructors or future readers can follow to understand the provenance of ideas.

Fostering Ethical AI Use

Just as students learn why plagiarism is wrong and how to avoid it, they now must learn what constitutes ethical vs. Unethical use of AI in academia. For example the State University of New York (SUNY) has planned to include instruction on the ethical dimensions of AI use as a requirement, ensuring students understand when and how AI can be used appropriately in academic work [18].

Teaching students to always disclose AI involvement is a fundamental outcome of such literacy: it trains them to treat AI like any other source or collaborator. Ultimately, normalizing the citation of AI supports a culture of openness.

How a Well-Balanced Solution Looks Like

While policy frameworks and guidelines set the stage, practical tools are needed to implement and secure process transparency at scale. One illustrative solution is **Mentafy** – a platform that embodies the balanced approach of combining process-based integrity transparency and pro-active student support. Though proprietary, its core principles can guide wider adoption globally.

The core idea behind Mentafy is *"evaluating solely after-the-fact does not work anymore – better analyze the process and help proactively"* in assessing academic work [8]. When a student authors a paper with Mentafy, the platform documents how the text was written - for example, recording when passages were typed, pasted, or possibly generated by an AI. Beyond the first simple percentage score, instructors can dive deep and review the authorship evidence in the student's text. Mentafy addresses plagiarism checker and AI-detector shortcomings, yielding the following benefits:

Higher Reliability

Mentafy analyzes actual writing behavior (e.g., rapid pasting vs. gradual revision), not just style or guesses, achieving greater accuracy than standard AI detectors. Accordingly, students can respond to specific evidence, and instructors can make informed decisions rather than trusting an algorithm blindly.
For instance, if a section of the essay appeared instantaneously and matches a known AI pattern, that's strong evidence of AI use. Conversely, if a student typed out a section slowly, over time, taking breaks, with revisions - that is rather evidence of original authorship.

Bias Mitigation

Mentafy examines writing patterns rather than language style, avoiding unfair bias against non-native speakers [19] by identifying behavior (e.g., copy-paste, rapid input) instead of vocabulary.

Embedding Education and Support

Most importantly, Mentafy is not about catching students - it is about guiding them. It provides writing support and real-time feedback about responsible AI use. For instance, if students paste large AI-generated sections, Mentafy encourages them to rewrite in their own voice. Mentafy CEO Markus Goldbach says, it helps students "acquire critical thinking and digital literacy alongside AI usage," clearly marking the boundaries of proper assistance.

Real-World Feasibility

Mentafy protects student privacy (no retention of deleted text, temporary storage only, no keylogging) and boosts instructor efficiency by automating analysis and flagging only suspicious cases.

"Integrity is doing the right thing, even when no one is watching", *C.S. Lewis*

Recommendation for Implementation

Transitioning to a process-transparency paradigm and away from pure detection requires coordinated changes in policy, practice, and tools. Below are key recommendations for educators, administrators, and policymakers to implement these ideas:

Update Academic Integrity Policies

Revise honor codes and integrity policies to explicitly address AI tools. Policies should define acceptable vs. unacceptable AI use for coursework and emphasize transparency. Include statements such as *"Students may use generative AI for preliminary research or drafting only if allowed by the instructor, and any AI assistance must be cited in the submitted work"*. Make it clear that undisclosed use of AI is considered plagiarism or cheating [15].

Integrate the AIAS Framework

Academic departments should consider adopting the Artificial Intelligence Assessment Scale (AIAS) when designing assignments and assessments [7]. This means deciding for each significant assignment which AIAS level (1 through 5) applies and communicating that to students in assignment instructions or syllabi. A writing-intensive course might mark some papers as "Level 1 – No AI allowed" to cultivate independent writing skills, while a capstone project could be "Level 4 – AI + Human Collaboration" to encourage innovative AI use with critical reflection.

Promote Responsible AI Use and Citation

Provide training and resources on how to use AI tools ethically and how to cite them. Institutions can develop quick reference guides for students on citation formats for AI (covering MLA, APA, etc. as relevant) [16] [17] and examples of what proper disclosure looks like (e.g., a student's note in an essay: "ChatGPT was used to help generate ideas for the structure of this paper, and one paragraph draft, which I then substantially revised."). Workshops or modules on digital literacy should include AI tool demos, showing both their useful capabilities and their limits or pitfalls.

Require Process Documentation

Implement requirements for students to submit evidence of their work process. This could include draft submissions, research logs, writing journals, or version histories – where possible, leverage technology (be cautious with AI detection tools [20]) to simplify this work for your staff and students. For written assignments, instructors might ask for one or two intermediate drafts or a brief reflection on how the students developed their essay. In courses where coding or data work is involved, students could submit code notebooks or revision histories. The key is to make it routine that a final submission is accompanied by a "behind-the-scenes" peek at how it came to be [4]

"Success Without Integrity Is Failure", ***Anonymous***

"35% of the students admit to cheat with AI", Lee et al. 2024

Leverage Technology (with Careful Rollout)

Consider adopting process-oriented integrity software to assist in monitoring and guiding student work. Begin with pilot programs to evaluate effectiveness and gather feedback from both faculty and students. It is crucial to involve stakeholders in selecting and configuring such tools – for example, ensuring that any data collected on student writing is secure and used solely for integrity purposes, to maintain trust. Provide training if you use technology, so instructors know how to interpret its reports and students understand how it functions. Emphasize that the tool is there not only to catch mistakes but to help students learn good academic practice – so what might be seen as a surveillance mechanism is understood as supportive educational aid.

Diversify Assessment Strategies

Reduce over-reliance on take-home written assignments as the sole measure of learning, especially in courses where AI could complete such tasks easily. Introduce a mix of assessment types that make misuse harder and engage students in varied ways. For example, incorporate more in-class writing assignments, oral presentations, group projects, and practical demonstrations of skills. Group work, if structured well, encourages peer accountability. Besides making it extremely difficult to outsource all this work to AI, these methods contribute to deeper learning and assess alternative skills.

Engage Students as Partners

Students are more likely to buy into policies that they feel have considered their perspective. Forming a committee or holding forums that include students can surface valuable insights – for instance, students might share which assignments tempt AI misuse and could be redesigned or spread the message of transparency among peers. When the student body understands that the institution's aim is to help them learn in a world with AI (and not to punish them arbitrarily), compliance with new measures will be higher.

Continuous Policy Review and Improvement

Regularly review the effectiveness of the implemented strategies – for example, track the number of academic misconduct cases related to AI before and after adopting these recommendations, gather feedback from faculty on whether student work quality or honesty has improved, and keep abreast of legal or ethical updates (data privacy laws, etc.). Be prepared to refine the AI usage guidelines and adopt new tools as they emerge.

Conclusion

The dawn of widespread generative AI in academia is not the end of academic integrity – but it does demand that we evolve our understanding and practices of integrity. As this whitepaper has argued, “beyond detection” lies a more resilient and educative approach: embedding transparency, guided AI usage, and honesty into the learning process itself.

Traditional detection-first strategies are neither reliable nor sufficient in the face of rapidly advancing AI capabilities. If we cling to them, we risk false accusations against students, arms-race dynamics, and a corrosive atmosphere of suspicion. Moreover, we miss the larger point: ensuring students learn and uphold values, rather than simply trying to catch them in wrongdoing. The example of Mentafy demonstrates that technology itself can be harnessed to uphold integrity constructively.

Process transparency offers a path forward. By requiring students to show their work and by integrating frameworks like AIAS, educators can deter misconduct by design and refocus on learning outcomes. Such measures, coupled with responsible AI citation practices, reinforce a culture of accountability and clarity. Students educated under these norms will understand that using AI is acceptable only if done openly and within set guidelines – a valuable ethical lesson for their future, where AI will be a common tool in many professions.

For policymakers and academic integrity officers, the charge is clear. We must update our policies and tools. As one expert provocatively asked, “Does our traditional definition of academic integrity still hold in the GenAI era?” [9]

The answer is that the core values – honesty, responsibility, fairness – remain, but the application must broaden. Honesty now includes admitting AI assistance; responsibility now includes learning to use AI properly; fairness now includes protecting students from unfounded AI accusations and providing equal access to AI literacy.

The answer is that the core values – honesty, responsibility, fairness – remain, but the application must broaden. Honesty now includes admitting AI assistance; responsibility now includes learning to use AI properly; fairness now includes protecting students from unfounded AI accusations and providing equal access to AI literacy.

References

[1] A. K. Kofinas, C. H. Tsay, and D. Pike, "The impact of generative AI on academic integrity of authentic assessments within a higher education context," *British Journal of Educational Technology*, Mar. 2025, doi: 10.1111/bjet.13585. Available: https://doi.org/10.1111/bjet.13585

[2] D. Weber-Wulff *et al.*, "Testing of detection tools for AI-generated text," *International Journal for Educational Integrity,* vol. 19, no. 1, Dec. 2023, doi: 10.1007/s40979-023-00146-z. Available: https://arxiv.org/abs/2306.15666

[3] S. Ankel, "A Texas professor failed more than half of his class after ChatGPT falsely claimed it wrote their papers," *Business Insider*, May 29, 2023. Available: https://www.businessinsider.com/professor-fails-students-after-chatgpt-falsely-said-it-wrote-papers-2023-5

[4] "Beyond AI detectors: Embracing Proof-of-Work to support academic integrity | PowerNotes." Available: https://www.powernotes.com/post/beyond-ai-detectors-embracing-proof-of-work-to-support-academic-integrity

[5] J. M. Lodge and The University of Queensland, "The evolving risk to academic integrity posed by generative artificial intelligence: Options for immediate action," 2024. Available: https://www.teqsa.gov.au/sites/default/files/2024-08/evolving-risk-to-academic-integrity-posed-by-generative-artificial-intelligence.pdf

[6] M. Anderson, "Detailed Guide for Citing ChatGPT," *PlagiarismSearch.com*, Mar. 17, 2025. Available: https://plagiarismsearch.com/blog/detailed-guide-for-citing

[7] "View of the Artificial Intelligence Assessment Scale (AIAS): A framework for ethical integration of generative AI in educational assessment." Available: https://open-publishing.org/journals/index.php/jutlp/article/view/810/769

[8] Meeraa, Jan. 06, 2025. Available: https://mentafy.com/2025/01/ai-driven-transparency-transforming-academic-integrity-with-mentafy/

[9] S. I. Dobrin, *A GUIDE FOR EDUCATORS Talking about Generative AI*. 2021. Available: files.broadviewpress.com/sites/uploads/sites/173/2023/05/Talking-about-Generative-AI-Sidney-I.-Dobrin-Version-1.0.pdf

[10] Reassessing Academic Integrity in the Age of AI: A Systematic Literature Review on AI and Academic Integrity," *Sciencedirect*. Available: https://www.sciencedirect.com/science/article/pii/S2590291125000269

[11] Cheating in the Age of Generative AI: A High school survey study of cheating behaviors before and after the release of ChatGPT," *Sciencedirect*. Available: https://www.sciencedirect.com/science/article/pii/S2666920X24000560

[12] C. Chaka and University of South Africa, "Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review," *Journal of Applied Learning & Teaching*, Feb. 2024, doi: 10.37074/jalt.2024.7.1.14.

[13] M. Perkins, J. Roe, D. Postma, J. McGaughran, and D. Hickerson, "Detection of GPT-4 generated text in higher Education: combining academic judgement and software to identify generative AI tool misuse," *Journal of Academic Ethics,* vol. 22, no. 1, pp. 89–113, Oct. 2023, doi: 10.1007/s10805-023-09492-6. Available: https://doi.org/10.1007/s10805-023-09492-6

[14] "How hard can it be? Testing the dependability of AI detection tools," *E Campus Learn, Share, Connect,* Oct. 15, 2024. Available: https://www.timeshighereducation.com/campus/how-hard-can-it-be-testing-dependability-ai-detection-tools

[15] "Academic integrity and syllabus support in the age of generative AI," *NYU Steinhardt*, Feb. 10, 2025. Available: https://steinhardt.nyu.edu/faculty-and-staff/academic-affairs/steinhardt-ai-hub/academic-integrity-and-syllabus-support-age

[16] "LibGuides: AI Literacy in the Age of ChatGPT: Citing generative AI." Available: https://libguides.library.arizona.edu/ai-literacy-instructors/transparency

[17] "Research Guides: Introduction to Academic Integrity: ChatGPT." Available: https://guides.lib.umich.edu/c.php?g=1039501&p=9763907

[18] "Chancellor King announces improvements to SUNY Undergraduate general education curriculum, including new focus on civic discourse," *SUNY*, Jan. 07, 2025. Available: https://www.suny.edu/suny-news/press-releases/1-25/1-7-25/general-education.html

[19] "GPT detectors are biased against non-native English writers." Available: https://arxiv.org/pdf/2304.02819

[20] L. Coffey, "Professors proceed with caution using AI-detection tools," *Inside Higher Ed* | Higher Education News, Events and Jobs, Feb. 09, 2024. Available: https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2024/02/09/professors-proceed-caution-using-ai