Beyond Detection: Rethinking AI Tool Use in Academic Theses

Sophie Schauer[1] and Katharina Simbeck[2]

[1]HTW Berlin University of Applied Sciences, Berlin, Germany,
sophie.schauer@htw-berlin.de
[2]HTW Berlin University of Applied Sciences, Berlin, Germany,
katharina.simbeck@htw-berlin.de

Abstract

Generative AI tools like ChatGPT are becoming an everyday practice in student writing, and academic institutions are under increasing pressure to respond. So far, most policies have focused on detection or prohibition. This position paper argues that detection and prohibition are not enforceable and calls for a shift toward a competence- and integrity-based framework. Using the case of the Business Computing program at HTW Berlin, we present a model policy that permits AI use under transparent conditions. Students are required to understand and defend their work. Further, the approach acknowledges that the use of AI reduces the skill set required to create academic texts. We thus argue that it is necessary to increase the level of difficulty in educational assignments and discuss opportunities to do so. Rather than treating AI as an external threat, this approach integrates it into academic practice. Future research will examine how such policies influence student performance and learning outcomes.

Keywords: *Generative AI in Education, AI Literacy, Higher Education Policy, AI Transparency.*

1. Introduction

In the past few years, generative AI tools have made their way into academic writing. For many students, these tools are now part of their everyday writing process. Tools are being used for brainstorming, structuring arguments or polishing their writing and checking grammar and spelling.

Academic writing skills have a long-established tradition in academic teaching. By explaining a concept in their own words, students not only gain but also demonstrate an understanding. By carefully selecting relevant and original references, students understand and prove how their work builds on scholarly tradition. By prioritising certain references and concepts, students intentionally select their approach to a subject. Academic writing is thus not only an assessment, but a major learning opportunity.

In the past years, many academic text-based professions have undergone significant change driven by generative AI. In text translation, production (journalism, copywriting) and editing efficiency have improved to an extent that a substantial reduction in jobs can be observed [1,2]. In the legal profession, AI is used to review contracts, assess claims and prepare briefs. In software engineering, AI-assisted code generation is integrated in standard workflows [3]. Higher education, thus, on the one hand, needs to acknowledge the prevalence of generative AI in the workplace and integrate AI literacy and data literacy into all curricula. On the other hand, education has always first taught the manual process (e.g. long addition) before allowing tool use (calculators). Of course, once tools are allowed to be used, the tasks change and level of difficulty increases. Pupils are not required to solve more and more advanced addition assignments using calculators, but they move on to new mathematical concepts.

Institutions have largely responded by trying to detect AI-generated content or ban its use altogether, often under the pretext of upholding academic integrity. But these measures raise more problems than they solve. Detection tools are unreliable and prone to false positives [4,5]. Building academic policies around their use is not only technically flawed but also pedagogically shortsighted and cannot be the foundation of academic integrity enforcement. Instead, we argue that higher education institutions need to rethink how they design assignments, assess student work and define acceptable use of AI tools. This shift is necessary if higher education wants to respond to the changes already happening in how students think, write and learn.

2. The Limits of Detection

Academics often believe that they can identify AI-generated texts submitted by students. Indeed, no study found that frequent LLM users can detect AI texts with greater accuracy, even without formal training. In practice, their judgement matched or surpassed the performance of automatic detectors. They base their judgement not only on surface-level features as detectors do but also on deeper cues such as originality and tone. Human detectors were found to be effective even against evasion tactics like prompt-based humanisation [6].

However, the limits of AI-detectors are well documented. The reliability for 14 detectors to classify human-written and AI-generated texts in an academic context was evaluated with less than 80% accuracy for most tools, and no tool showed consistent reliability across all text types [8]. False negatives were common, with some tools missing over 50% of the cases where AI text was paraphrased or manually edited. Most of the investigated tools showed a bias toward classifying text as human-written. From all 14 tools, Turnitin ranked highest but still misclassified a significant number of paraphrased or edited texts. Tools like Content at Scale failed almost completely and classified nearly all texts as human-written [8]. In a more recent study, three detectors were systematically tested [4]. LLM prompts were modified to produce outputs that could not be detected by popular detection tools. Basic AI-texts are usually reliably detected. However, when prompts deliberately include imperfections such as minor grammar errors, inconsistent sentence structure and less polished wording, the detection rates drop [4]. Adding stylistic texts that had a lower readability score was more often seen as human. In contrast, polished academic-sounding texts were typically flagged as AI. All three tested detectors showed vulnerabilities to prompt manipulation [4]. Another similar study found ZeroGPT to have high false positive and false negative rates as well as a significant increase in false negative rate when using ChatGPT 3.5 to paraphrase the original AI text [7].

Another concern is false positives, where detectors often misclassify writing by non-native English speakers that use translation tools or texts with repetitive academic phrasing as AI-work [5,8]. Another core issue is opacity, as most detectors do not disclose how they classify texts, what features they weigh most heavily or how they handle edge cases. Their scores and decisions offer little transparency and accountability.

In summary, even though educators believe that AI texts can be detected, there is no technically reliable way to do so if students use specific prompting strategies or manually adapt the generated text. The use of AI detectors thus only serves to classify students by AI literacy: students with advanced prompting techniques cannot be identified.

3. Proposed AI Policy for Business Computing Theses

The department of Business Computing at HTW Berlin has adopted structured guidelines that acknowledge AI's presence in the academic process but set firm boundaries for transparency and fairness. These regulations are valid from the winter semester 2024/25 onward and will be reviewed continuously with the development of the field. The objective of the policy is to create clear rules for students that are enforceable and that do not disadvantage honest students.

First and foremost, the use of AI tools is explicitly permitted. However, every thesis must include a detailed AI directory listing all AI tools used, specifying where and how they were applied (see Table 1 for an example). The requirement ensures that supervisors and examiners can understand the extent and context of AI involvement. Students are not required to submit a list of prompts used or provide a chat history of the conversation with the AI as this would not be enforceable.

AI Tool	Part of the Thesis (Page)	Purpose of Use
DeepL	Abstract (p. I)	Translating the German abstract into English, for rough structure and ideas. Final translation in own words.

ChatGPT	Section 3.3.4 Linguistic Variation in Input (p. 51 ff.)	Generating synonyms, similar phrasing, and correcting typos.
ChatGPT	Section 6 Implementation (p. 74 ff.)	Checking code for errors during development.
ChatGPT	Section 6.2 Database Methods (p. 77 ff.)	Support in generating SQL code, adding comments, and storing results in Python.
ChatGPT	Appendix A1: Bot Intents (p. XIV)	Extracting intents from Excel table and creating a simple listing for the thesis.

Table 1 Exemplary AI listing

A central part of this policy is the student's own understanding. All content submitted must be fully understood by the student. During the oral defence, students need to be able to answer questions about the text submitted by them. They are expected to articulate their reasoning, justify their choices and demonstrate a familiarity with all sources and results. Inability to do so is treated not as a misunderstanding but as an attempt at deception.

Equally important is the requirement of a structured and traceable scientific methodology. The process must be documented and defensible in discussion. AI tools may assist, but critical thinking and methodical reasoning remain non-negotiable.

This model creates a space for responsible AI use while placing accountability on the students. In doing so, it offers a practical way forward for higher education institutions.

4. Redesigning Tasks and Assessment Criteria

Given the availability of powerful tools, the structure of academic assignments must evolve. Attempting to uphold traditional formats while banning or ignoring AI use only creates uncertainty for students and inconsistent enforcement for educators. We therefore call for the adoption of different and more challenging assignment tasks. In parallel, the assessment criteria need to be updated and less emphasis needs to be put on aspects of academic work that AI can support, such as the related work section of a thesis project. Rather than simply assessing the final outputs, educators should pay closer attention to how students arrive at their conclusion. Structuring arguments, interpreting data or weighing conflicting perspectives are areas where human judgement is irreplaceable.

Academic thesis topics and evaluation criteria must be adapted to the availability of generative AI. In business computing, tasks must put greater emphasis on analysis of real-world business processes, implementation of prototypes, building of machine learning models on datasets, collection and analysis of own data or contribution to open source projects. Students' own contributions must go beyond review of literature.

Evaluation of academic theses requires adaptation as well. In the past, poor style often correlated with superficial content. Now, even low-skilled students can generate polished-sounding text fragments, and educators need to invest more time into reading and assessing text. The integration of the students' own work into the state of the art needs to be given higher emphasis, while the related work section cannot be considered deciding anymore.

Table 2 provides an overview of typical sections or tasks in a thesis in a business computing thesis, how generative AI impacts the level of difficulty of those sections and which consequences could be drawn by evaluators with regard to task design and assessment criteria.

Task	Impact of Generative AI	Consequences for Task Design and Assessment
Write related work section	Theoretical sections of thesis can easily be generated including correct references.	▫ Less focus on explanation of basic concepts. ▫ Stronger focus on selection of specific concepts relevant for topic at hand and for integration of concepts identified and explained into methodology (example: is questionnaire built reflecting concepts discussed in related work). ▫ Strongly reduced weight of related work section in the assessment.
Describe and document business process	Little impact for description of real process with partner company.	▫ Preference for real world tasks as opposed to theoretical topics. ▫ Emphasize that student must document approach and process of information collection in company. ▫ Student must identify specific risks and opportunities. ▫ Increased weight in assessment.
Prototypical implementation (software components, databases), data analysis	Implementation significantly enhanced by generative AI.	▫ Students can be given much more challenging tasks. ▫ Emphasize that student must document implementation process (definition of requirements, implementation concept and technological choices made, testing). ▫ Increased weight in assessment.
Empirical data collection	Questionnaire design or similar activities enhanced by generative AI.	▫ Students are expected to ground their empirical approach in theoretical understanding. ▫ Emphasize that student must document approach and process of information collection in company. ▫ Increased weight of methodological rigor in assessment (example: student successfully recruited a large number of high profile participants for explorative interviews).

Documentation of code and data	General need to increase level of difficulty in thesis projects.	Students are required to document code and data according to FAIR research data management practices to ensure scientific transparency and reproducibility. Both code and data must be made available to reviewers with a readme file that specifies type and structure of code and data as well as instruction on how to use it. Code must be provided with clear instruction on how it needs to be run (e.g. required libraries, which file to run first). Data must be provided with information on how data was collected and analysed, how consent was obtained (if applicable). Students are required to explicitely select a license for code and data. This is a new part of assessment criteria.
Applying appropriate academic style	Writing correct and stylistically appropriate text significantly enhanced by generative AI.	▫ Lack of academic rigor cannot be easily identified by evaluators based on style anymore. ▫ Students are expected to use generative AI to improve style.

Table 2 Typical tasks in Business Computing thesis projects and impact of generative AI

In summary, redesigning assessment is not about lowering standards, but about aligning them with the changing reality. In fact, expectations may rise, as students are allowed to use these tools and can therefore produce higher-quality output and gain deeper insights.

5. Conclusion and Outlook

Educators need to accept the widespread use of generative AI in academic writing as a reality. Relying on detection is both technically unreliable and pedagogically limiting. As we argued in this paper, detection tools cannot offer the reliability or fairness needed to support academic integrity practices. Generative AI challenges century-long traditions in academics. Institutions need to develop clear and realistic norms for its use. This shift demands a rethinking of how learning tasks are designed and how student work is evaluated. The next step will be to examine how such implemented frameworks shape student engagement and learning outcomes.

At the same time, we acknowledge the need for writing tasks that are not AI-supported and which will consequently need to be conducted in a supervised classroom setting. Finally, we would like to draw the reader's interest to the fact that written student work cannot only be generated using AI but that generative AI is a powerful tool in helping to assess written assignments. A dystopic educational future where students' AI-generated texts are automatically assessed by AI tutoring systems can easily be imagined. The future discussion on AI policies in education should thus include both the student's and the educator's perspective.

6. Acknowledgements

This paper portrays the work carried out in the context of the EUonAIR project (101177370) generously funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

7. References

[1] David H., Frank Levy, and Richard J. Murnane. "The skill content of recent technological change: An empirical exploration." The Quarterly journal of economics 118, no. 4 (2003): 1279-1333.

[2] Qiao, Dandan, Huaxia Rui, and Qian Xiong. "AI and jobs: Has the inflection point arrived? Evidence from an online labor platform." arXiv preprint arXiv:2312.04180 (2023).

[3] Sergeyuk, Agnia, Yaroslav Golubev, Timofey Bryksin, and Iftekhar Ahmed. "Using AI-based coding assistants in practice: State of affairs, perceptions, and ways forward." Information and Software Technology 178 (2025): 107610.

[4] Ernst, Marie, Fabian Rupp, and Katharina Simbeck. "You Can't Detect Me! Using Prompt Engineering to Generate Undetectable Student Answers." Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025), no. 2 (2025).

[5] Liang, Weixin, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. "GPT detectors are biased against non-native English writers." Patterns 4, no. 7 (2023).

[6] Russell, Jenna, Marzena Karpinska, and Mohit Iyyer. "People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text." arXiv preprint arXiv:2501.15654 (2025).

[7] Weichert, James, and Chinecherem Dimobi. "Dupe: Detection undermining via prompt engineering for deepfake text." arXiv preprint arXiv:2404.11408 (2024).

[8] Weber-Wulff, Debora, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. "Testing of detection tools for AI-generated text." International Journal for Educational Integrity 19, no. 1 (2023): 1-39.