From NLG Evaluation to Modern Student Assessment in the Era of ChatGPT: The Great Misalignment Problem and Pedagogical Multi-Factor Assessment (P-MFA)

Mika Hämäläinen and Kimmo Leiviskä
Metropolia University of Applied Sciences
Helsinki, Finland
first.last@metropolia.fi

Abstract

This paper explores the growing epistemic parallel between NLG evaluation and grading of students in a Finnish University. We argue that both domains are experiencing a Great Misalignment Problem. As students increasingly use tools like ChatGPT to produce sophisticated outputs, traditional assessment methods that focus on final products rather than learning processes have lost their validity. To address this, we introduce the Pedagogical Multi-Factor Assessment (P-MFA) model, a process-based, multi-evidence framework inspired by the logic of multi-factor authentication.

1 Introduction

In recent years, both the field of computational creativity and university pedagogy have faced a growing crisis of evaluation. In creative natural language generation research, Hämäläinen and Alnajjar (2021a) articulated the Great Misalignment Problem, which is the disconnection between a system's problem definition, its implemented method and the evaluation criteria used to assess its performance. This misalignment leads to superficial or misleading conclusions about a system's success, as the evaluation often fails to measure what the system was designed to achieve.

A similar issue now permeates higher education: the act of grading has become increasingly detached from the authentic learning processes it is meant to assess. As generative models such as ChatGPT[1] can produce fluent, original-looking outputs on demand, educators can no longer be sure whether a student's submitted work reflects genuine understanding or merely the clever use of an external artificially cognitive tool.

Just as computational creativity systems generate artefacts whose internal reasoning is opaque, students in the AI-saturated learning environment can now present creative products without revealing the intellectual pathway that led to them. In both cases, the evaluator encounters an artefact, such as a poem, a program, an essay or a design, without direct access to the underlying process that produced it. The traditional product-based evaluation paradigm, which assumes a transparent correspondence between output and competence, thus collapses. The opacity of generative systems mirrors the opacity of modern student work: both appear impressive on the surface, yet the link between creation and creator, between performance and understanding, is obscured.

This parallel suggests that the Great Misalignment Problem has re-emerged in pedagogy under new conditions. In education, the "problem definition" corresponds to the intended learning outcomes; the "method" is the student's learning process; and the "evaluation" is grading. When these three components fall out of alignment - when grades reflect polished submissions rather than genuine cognitive engagement - the integrity of learning assessment is compromised. The university thus faces a dilemma akin to that of computational creativity research: it risks optimizing for the appearance of creativity and competence rather than for the authentic development of these qualities.

To resolve this, both AI research and pedagogy must shift their evaluative focus from product to process. In computational creativity, meaningful evaluation requires attention to how a system produces its output, its generative mechanisms, constraints, and reasoning. Likewise, effective pedagogy must emphasize the learning process itself: reflection, iteration, collaboration, and the student's evolving relationship with knowledge. In both domains, understanding the process behind the product restores transparency, accountability, and interpretability. The challenge, then, is not merely to detect misuse of generative tools but to redesign evaluation frameworks that value the act of cre-

[1] https://chatgpt.com

Proceedings of the 10th International Workshop on Computational Linguistics for Uralic Languages, pages 1–5

ation—the unfolding of thought—as the true site of learning and creativity.

2 Related Work

The emergence of LLMs has received a mixed response among educators; several feel that AI is a threat to students' learning (Yu, 2023; Lin et al., 2023; Ogugua et al., 2023), while others focus on it's potential in future of education (Lo, 2023; Cleland Silva and Hämäläinen, 2024; Morgan, 2024; Macias et al., 2024).

Evaluation of NLG systems, creative and regular, has received a fair share of attention in the past (Clark et al., 2021; Freitag et al., 2021; Howcroft et al., 2020). Some researches even highlight that automated evaluation methods such as BLEU are simply not sufficient (Reiter, 2018). Evaluation mainly relies on evaluating the output of such systems rather than taking the creative process into account (Hämäläinen and Alnajjar, 2021b).

In terms of pedagogy, how people learn and how they should be taught is a relatively well-understood phenomenon. Frameworks such as Bloom's (1956) taxonomy, deep learning (see McGregor 2020) and constructive alignment (Biggs, 1996) have been widely used. However, there's a big gap between theory and practice, and ultimately, student assessment relies on grading some sort of a final product of learning such as an essay or thesis.

3 The Great Misalignment Problem in NLG Evaluation

In their work, Hämäläinen and Alnajjar (2021a) identify what they term *the Great Misalignment Problem* in the context of human evaluation in natural language generation (NLG). The core claim is that in much of NLG research that relies on human judgments, there is a systematic misalignment among three key components: (1) how the research problem is defined, (2) how the proposed method or model is formulated and (3) how the human evaluation is conducted. When these three are not tightly aligned, the authors argue, the validity, interpretability and reproducibility of human evaluation outcomes are severely compromised.

The authors support their claim by surveying ten randomly selected papers from ACL 2020 that include human evaluation. In their analysis, they examine (a) whether the problem definition is clearly and narrowly stated, (b) whether the proposed method follows directly from that problem definition, and (c) whether the human evaluation aligns both with the definition and with what the method is intended to model. They report that in only a single case among the ten did all three align. In many cases, the evaluation either ignores aspects of what the method is modeling or tests orthogonal criteria not grounded in the stated problem.

The implications of the Great Misalignment Problem are profound. Because human evaluation may end up measuring something other than what the model is intended to do, the reported improvements or differences in scores cannot reliably be attributed to the proposed method. Instead, they might arise from unintended artifacts, evaluation design biases, evaluator variance or other uncontrolled factors. This undermines claims of advancement, makes comparison across systems less meaningful, and complicates reproducibility. Moreover, the authors point out that human evaluation in NLP is often conducted with insufficient methodological rigor (e.g. vague questions, low numbers of judges and opaque protocols), further exacerbating the misalignment.

To move forward, the authors recommend that NLP researchers take the problem definition seriously and design methods and evaluations so as to maintain alignment. Concretely, they urge narrowing broad, vague problem statements into more precise, measurable sub-tasks; ensuring that modeling decisions correspond to those sub-dimensions; and crafting evaluation questions that directly probe the modeled behavior. They also call for full transparency in evaluation setup (e.g. prompt wording, judge selection and instructions) and suggest that human evaluation practices in NLP could benefit from importation of best practices from fields accustomed to subjective measurement (e.g. social sciences). They do not advocate abandoning human evaluation altogether, but rather reforming it so that it becomes a more trustworthy and interpretable component of NLP research.

4 LLMs and Teachers' Nightmare

Not unlike the findings described by Hämäläinen (2024), we have encountered negative teacher narratives in Metropolia University of Applied Sciences regarding LLM tools. There is a lot of fear among teachers that students would use the new technology to cheat and ultimately pass their courses without learning much. On the other hand, there are

also teachers who embrace the new technology and actively use it in their teaching.

It is fair to say that LLMs have introduced a radical disruption to the long-standing epistemic contract between teachers and students. Pedagogically, this contract rests on an implicit trust: that the student's submitted work represents their own intellectual labor and engagement with the learning process. The teacher, in turn, evaluates this work as evidence of learning, understanding and skill development. However, with the advent of LLMs capable of producing contextually appropriate, grammatically flawless and even stylistically distinct texts, this foundational trust is breaking down. Teachers now face the uneasy possibility that a student's polished essay or thoughtful reflection may be more a testament to prompt engineering than to actual comprehension.

From a pedagogical standpoint, this collapse of evaluative certainty threatens the very rationale of assessment (see Şahin et al. 2024; Fagbohun et al. 2024). If an assignment can be completed without genuine learning, then grades cease to measure educational achievement. They become measures of access to tools and skill in their manipulation. The anxiety many teachers experience arises not merely from the fear of academic dishonesty but from the erosion of pedagogical meaning itself. When outputs can no longer be reliably linked to the cognitive processes they are meant to demonstrate, the educational system loses its anchor: learning becomes performative rather than transformative.

The teacher's nightmare is not that students are cheating, but that the act of evaluation has become epistemically hollow. Traditional assessment methods such as essays, reports and even project work are predicated on a production model of learning where outputs reflect mental effort. In the LLM era, this assumption no longer holds. The teacher cannot see how the student arrived at a conclusion, whether the reasoning was genuine or algorithmically scaffolded. Consequently, feedback loses precision, as it targets the product rather than the learner's cognitive or creative process. Pedagogically, this creates a feedback loop of disengagement: students learn to outsource tasks, and teachers, sensing the futility of assessment, may lower expectations or turn toward increasingly mechanistic forms of surveillance.

In essence, LLMs expose the brittleness of output-oriented pedagogy. The crisis they introduce is not technological but epistemological: education must now grapple with redefining what it means to know, learn, and create in a world where human and machine outputs are indistinguishable. Teachers' frustration, then, is not only emotional but structural. It stems from a system designed for a pre-AI understanding of authorship and agency. The nightmare can only end when pedagogical practices evolve to prioritize the evaluation of the learning process itself over the evaluation of a final product, reclaiming assessment as a shared inquiry into learning rather than a judgment of finished artefacts.

The Great Misalignment Problem offers a valuable lens for rethinking student assessment. By highlighting the dangers of disconnect between problem definition, method and evaluation, the same critique encourages educators to scrutinize whether grading practices truly measure the intended learning outcomes. **In pedagogy, this means aligning what we want students to learn (problem definition), how they engage in that learning (method), and how we evaluate their understanding (evaluation**).

Rather than focusing solely on the final product, educators can design assessments that make the learning process visible through reflective writing, process logs, peer discussions or iterative project development. In doing so, the evaluation becomes an inquiry into alignment itself: does the student's process reflect the intended learning goals, and does the teacher's evaluation capture that process accurately? This alignment-centered approach transforms grading from an act of judgment into an act of dialogue, ensuring that assessment remains meaningful even in an era when the boundary between human and machine creativity is increasingly blurred.

5 Future of Grading: a P-MFA Approach

We propose a novel framework named Pedagogical Multi-Factor Assessment (P-MFA) for grading and learning evaluation designed for the age of generative AI. It builds on the logic of multi-factor authentication (see Ometov et al. 2018): just as digital security no longer relies on a single password, educational assessment should not depend on a single artefact such as an exam or essay. P-MFA therefore verifies learning through multiple complementary "factors," each representing a distinct dimension of competence—what the student knows (knowledge), produces (outputs), can do (ap-

plication), sustains over time (process continuity), reflects upon (self-evaluation), and connects to real contexts (situated understanding).

By combining these factors, teachers and students co-construct a trustworthy, multi-channel record of learning that is transparent, individualized, and resistant to the misuse of AI. Rather than focusing on control or detection, P-MFA shifts assessment toward alignment: ensuring that what is defined as learning, practiced as learning, and evaluated as learning all converge in an authentic and human-centered educational process.

Theoretically, P-MFA can be understood as a synthesis of constructive alignment and the Great Misalignment Problem. Constructive alignment posits that meaningful learning occurs when intended learning outcomes, teaching activities, and assessment tasks are coherently designed to support one another. The Great Misalignment Problem, in contrast, diagnoses the breakdown of such coherence in research evaluation: when problem definition, method, and evaluation diverge, the resulting claims lose validity.

P-MFA translates this alignment imperative into pedagogy for the generative-AI era, explicitly designing assessment systems that keep the "problem definition" (learning outcomes), the "method" (student learning processes), and the "evaluation" (grading practices) in continuous dialogue. Where constructive alignment emphasizes curriculum design, P-MFA operationalizes alignment through evidence diversity: multiple, process-anchored factors that ensure the assessment remains faithful to both the intention and the practice of learning. In doing so, P-MFA not only safeguards educational integrity against AI-generated artefacts but also reframes assessment as an interpretive act of maintaining epistemic alignment between what learning is meant to achieve, how it unfolds, and how it is ultimately recognized.

In essence, the P-MFA approach operationalizes the Great Misalignment Problem's philosophical insight within the classroom. It demands that educators explicitly design for a**lignment across definition, method, and evaluation**, ensuring that the assessment of learning remains meaningful, transparent, and resilient to technological disruption. By requiring multiple, process-anchored proofs of understanding, P-MFA not only protects the integrity of grading but also redefines it as a dynamic act of alignment, in which learning and evaluation evolve together.

Problem definition in P-MFA is reframed as the articulation of learning outcomes that go beyond static knowledge. Education's aim is no longer to verify that students can produce isolated outputs but that they can understand, apply, reflect and contextualize their knowledge

The Method of P-MFA corresponds to the pedagogical and learning practices through which these factors are activated. Instead of viewing learning as a linear input–output pipeline, P-MFA promotes iterative, reflective and contextual engagement. Students are not passive respondents to tasks but co-designers of their assessment trajectory, selecting factors that align with their goals and contexts

Finally, Evaluation in P-MFA is no longer an isolated measurement but a process of triangulation. Each factor functions as an evaluative lens that confirms or challenges the authenticity of others. The resulting alignment between what was meant to be learned, how learning occurred, and how it is assessed embodies the very correction the Great Misalignment Problem paper sought in NLP research. When teachers adopt a P-MFA framework, grading transforms from a verdict into an inquiry: a structured investigation into whether the student's demonstrated process and outputs align with the course's learning definition. This ensures that evaluation measures authentic engagement rather than algorithmic fluency. Moreover, because AI cannot convincingly reproduce personal reflection or contextual relevance, P-MFA restores pedagogical trust by embedding evaluation in dimensions that remain uniquely human.

6 Conclusions

The challenges faced in both computational creativity and contemporary pedagogy converge on a single epistemic issue: the difficulty of evaluating outputs without understanding the processes that produced them. The Great Misalignment Problem revealed how research can lose validity when problem definition, method, and evaluation drift apart, an insight that now illuminates the crisis of grading in the era of generative AI. Our Pedagogical Multi-Factor Assessment (P-MFA) model offers a concrete response to this challenge by embedding assessment within the learning process itself. Through its multi-factor design—combining evidence of knowledge, production, application, continuity, reflection, and context—P-MFA restores alignment between what learning is intended to

achieve, how it unfolds, and how it is recognized. In doing so, it reclaims evaluation as a transparent and dialogic practice, reaffirming the role of assessment not as an act of surveillance or verification, but as an interpretive inquiry into human understanding and growth in an age increasingly mediated by machines.

References

John Biggs. 1996. Enhancing teaching through constructive alignment. *Higher education*, 32(3):347–364.

Benjamin S Bloom. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Tricia Cleland Silva and Mika Hämäläinen. 2024. Innovating for the future: Ai and hrm capabilities for sustainability in higher education. In *Academy of Management Annual Meeting*, volume 2024.

Oluwole Fagbohun, Nwaamaka Pearl Iduwe, Mustapha Abdullahi, Adeseye Ifaturoti, and OM Nwanna. 2024. Beyond traditional assessment: Exploring the impact of large language models on grading practices. *Journal of Artificial Intelligence, Machine Learning and Data Science*, 2(1):1–8.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Mika Hämäläinen. 2024. Legal and ethical considerations that hinder the use of LLMs in a Finnish institution of higher education. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 24–27, Torino, Italia. ELRA and ICCL.

Mika Hämäläinen and Khalid Alnajjar. 2021a. The great misalignment problem in human evaluation of NLP methods. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 69–74, Online. Association for Computational Linguistics.

Mika Hämäläinen and Khalid Alnajjar. 2021b. Human evaluation of creative NLG systems: An interdisciplinary survey on recent papers. In *Proceedings of the First Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 84–95, Online. Association for Computational Linguistics.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Shu-Min Lin, Hsin-Hsuan Chung, Fu-Ling Chung, and Yu-Ju Lan. 2023. Concerns about using chatgpt in education. In *International conference on innovative technologies and learning*, pages 37–49. Springer.

Chung Kwan Lo. 2023. What is the impact of chatgpt on education? a rapid review of the literature. *Education sciences*, 13(4):410.

Melany Vanessa Macias, Lev Kharlashkin, Leo Einari Huovinen, and Mika Hämäläinen. 2024. Empowering teachers with usability-oriented llm-based tools for digital pedagogy. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 549–557.

Sue LT McGregor. 2020. Emerging from the deep: Complexity, emergent pedagogy and deep learning. *Northeast Journal of Complex Systems (NEJCS)*, 2(1):2.

Hani Morgan. 2024. Implementing chatgpt to support teachers and promote learning. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, 97(5):125–133.

Divine Ogugua, Seong No Yoon, and DonHee Lee. 2023. Academic integrity in a digital era: Should the use of chatgpt be banned in schools? *Global Business & Finance Review (GBFR)*, 28(7):1–10.

Aleksandr Ometov, Sergey Bezzateev, Niko Mäkitalo, Sergey Andreev, Tommi Mikkonen, and Yevgeni Koucheryavy. 2018. Multi-factor authentication: A survey. *Cryptography*, 2(1):1.

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Alper Şahin, Nathan Thompson, and Kadriye Ercikan. 2024. Opportunities and challenges of ai in educational assessment. *Journal of Measurement and Evaluation in Education and Psychology*, 15(Special Issue):260–262.

Hao Yu. 2023. Reflection on whether chat gpt should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology*, 14:1181712.