doi:10.2533/chimia.2025.684 Chimia 79 (2025) 684–688 © G. Kortemeyer

Have to Shake It Up: STEM Education Feeling the Heat from Artificial Intelligence

Gerd Kortemeyer*

Abstract: General-purpose AI already correctly solves most traditional assessment problems in first-year STEM education, and it continues to become more proficient with every release. This quiet superheating of familiar practices by increasing AI capabilities will yield a messy eruption unless instructors introduce deliberate 'nucleation sites' and shake up the curriculum. We argue that results of science education research are now more relevant than ever since they provide insights and strategies on how to optimize human learning. Finally, we emphasize the importance of community and wellbeing – through peer instruction, studios, and brief, structured oral checks with whiteboarding – to counter loneliness and fatalism. As most of these instructional methods involve open-ended tasks, which tend to be more resource-intensive in grading and feedback than traditional, solution-oriented closed-form answer practices, AI can play an important role in assisting and supporting human educators. Thus, we illustrate how ETH Zurich's *Ethel* system operationalizes these approaches through a course-grounded chatbot, formative feedback on handwritten work, on-demand practice, and grading assistance, while keeping learning human.

Keywords: Artificial intelligence · Assessment · Chatbots · General chemistry

1. Introduction

General-purpose AI has crossed meaningful thresholds for first-year university sciences. In mathematics, physics, computer science, and chemistry, models can now solve a wide range of routine tasks, draft plausible reasoning, and even grade parts of handwritten exams with surprising fidelity. Like a test tube held motionlessly over a Bunsen burner, our teaching and assessment practices risk bumping: quiet superheating followed by a sudden, messy eruption, if we ignore these constantly increasing capabilities. To prevent this, it is metaphorically required to agitate the tube or add boiling chips: small, deliberate interventions that create nucleation sites. An analogous set of interventions may be needed in chemistry education: shake up how we teach, what we ask, how we ask it, and how we grade it, so we channel AI's heat into productive learning.

2. Where AI Already Runs Hot, and Where it Still Sputters

In 2023, GPT-3.5 could barely pass an introductory physics course, making algebraic errors and reasoning worse than the average student;[1] graphics back then were not even an option. With every new version, generally the models acquired more and more STEM capabilities.

An illustrative example is a General Chemistry exam given at our institution: a mix of multiple-choice, drawing, graph interpretation, stoichiometry, numerical calculations, and short-answer questions. GPT-4o, the most recent version of the OpenAI models in 2024, was prompted to solve this exam and received 46.9/60 points – about as good as the best students in the course. Doing the same task with a May-2025 model, simply based on the uploaded exam PDF-document with no prompting beyond 'solve this exam', GPT-5-Thinking scored 55/60 – coming out better than any student in the course.

While the increase of 13.5 percentage points in one year might not be that impressive, what is remarkable is the improvement in capabilities: for example, GPT-4o predicted that the pH value of a solution of sodium hydrogensalicylate (c = 0.0025 M) should be slightly basic, which is correct. The subsequent calculation was, however, wrong, as the pH value was determined to be 5.5 (instead of 7.1). The system had neither the numerical ability to correctly calculate the value nor the reasoning ability to catch that the result would be acidic, contradicting the original prediction. The system also failed to balance a chemical equation corresponding to a redox reaction or did not manage to properly determine the final temperature of a system consisting of 3 moles of liquid water H_2O at a temperature of 100 °C and 1 mole of ice H_2O at a temperature of 0 °C (73.6 °C instead of 55 °C) – likely 'forgetting' to heat the melted ice from 0 °C to the final temperature. Finally,

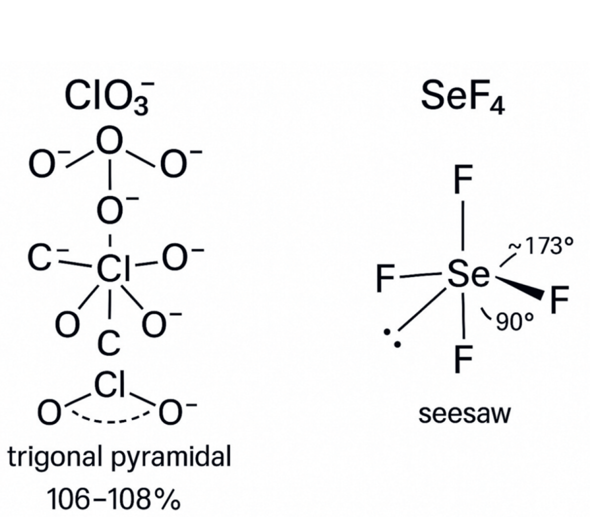

Fig. 1. AI-produced sketch for drawing Lewis structures of a chlorate ion and selenium tetrafluoride; still some room for improvement here.

Correspondence: Dr. G. Kortemeyer, E-mail: kgerd@ethz.ch
ORCID: 0000-0001-6643-9428.
Rectorate and ETH AI Center, ETH Zürich.

issues arose on graphical items: describing a line in a diagram and drawing structures. A year later, only the problems with graphics remained (see Fig. 1).

Chemistry is not the only example; similar results exist in physics: strong on concepts,[2] problem-solving, symbolic and numerical exercises, as well as narrative tasks,[3] still weak on graphical tasks.[4] How long will it take until generative AI also masters those? Maybe that question is irrelevant for us as educators: we should simply work under the assumption that it will eventually score 60/60 and start considering what that means.

3. When AI is Performing Better than Any of Our Students

As science educators, we have a mostly accepted set of learning objectives for introductory courses, and we assess if those goals have been reached with mostly accepted instruments. In general chemistry, this would be the familiar though sometimes disputed arc of atomic structure and periodic trends; bonding and molecular structure; stoichiometry and gases; thermodynamics, equilibria, acids–bases, kinetics; and the laboratory practices that support them.[5,6] We say we want students to build durable conceptual understanding, to move fluently among macroscopic, particulate, and symbolic representations, to estimate and check reasonableness, to model and justify assumptions, to argue from data, and, ideally, to develop some curiosity and enjoyment of chemical thinking.[7] Yet we typically certify this learning with timed, highly structured assessments: algorithmic problem sets, short-answer or multiple-choice exams keyed to canonical procedures, brief lab write-ups graded for compliance. Those instruments are reliable and convenient, but they often reward speed, pattern-matching, and formatting answers more than sense-making, explanation, or uncertainty management. In effect, we treat success on closed-form proxies as evidence of the richer outcomes we care about, while our 'hidden curriculum' remains tacit.[8] Now AI solves all those closed-form assessment instruments. But AI does not experience curiosity or develop an epistemology; it can simulate expert-like responses. The disconnect becomes more and more blatant, and if we do not do anything, humans will not develop those things either if we keep teaching and assessing like we do.

As AI cranks up the heat on those proxy instruments and makes them increasingly meaningless, we need to shake up the traditional canon of instruction and assessment, or eventually the construct is going to pop. Students in introductory courses, particularly non-majors, will ask themselves why they have to do those tasks if they can never perform better than AI. Not seeing the value, traditional homework exercises may easily degenerate to exercises where students take photos of the assignments, upload them, and prompt an AI to 'solve this.'[9]

4. Shaking, Not Simmering: Making AI Part of the Workflow Without Letting It Define Learning

We do not need a revolution, and certainly not 'AI everywhere' with miracle expectations. Generative AI is a reality that will unlikely go away, and we need to consider it in our teaching, but human learning has aims and mechanisms that are not machine learning; the vision of students learning alone from a [*the reader may insert a series of hype phrases*] AI is, at best, misguided. The productive move in an AI-rich era is to double down on evidence-based practices from science-education research, namely active engagement, feedback-rich tasks, explanation, and uncertainty work, while using AI selectively to reduce their resource burden (*e.g.* generating practice and formative feedback, curating materials, easing logistics), not to replace pedagogy. At ETH Zurich, we are continuously building an AI-ecosystem called *Ethel* to assist the implementation of proven, research-based educational strategies.[10]

5. Why a 'Process-First' Emphasis Works

Undergraduate science education is most effective when students express how they think, not only what they get. Decades of research show that self-explanation, planning before calculation, explicit statement of assumptions, and routine checks (units, scale, limiting cases, uncertainty) strengthen transfer, reduce careless error, and make misconceptions visible.[11] Students need to submit a 'trace' with every substantial problem, and the rubric assigns weight to plan quality, representation choices, and post-hoc checks alongside correctness.

Ethel supports this stance without doing the thinking for students. When a student asks the course chatbot for help on a calculation, in several courses, it first returns a plan skeleton and points to the relevant page of the course notes rather than a finished answer; when students upload handwritten work for formative feedback, the system highlights missing assumptions or absent unit/scale reasoning – this mechanism closes the feedback loop on formative assessment in a timely, always-available fashion without being punitive. Of course, sometimes the algorithms might not be able to decipher the handwriting or still make logical mistakes – but students tell us that this is okay, and 'fighting' with the AI helps them learn. Students at technical universities are not naïve and do not blindly believe in AI statements, so we believe that in these particular formative phases of learning, we did not pre-verify every AI comment during formative use; students were encouraged to challenge feedback and verify against course notes.

6. Tasks that Resist Shallow Shortcuts

The most authentic chemistry tasks ask students to model the phenomena in order to design, justify, and critique.[12] This is work that requires crossing among macroscopic phenomena, particulate models, and symbolic representations. Messy experimental data, raw spectra with artifacts, mechanism arguments that demand arrow pushing under stereochemical constraints, and critiques of plausible but flawed explanations all create 'nucleation sites' for learning. They reward warrants and trade-offs, not pattern-matching.

Here the chatbot's course-specific retrieval is an asset: instead of generic web summaries, it points students back into the instructors' slides and scripts to justify model choices or parameter settings.

Of course, not all exercises can be on these deep analysis levels. Early in the learning process, there is still a place for simpler application, comprehension, and even knowledge questions. Generating these can be left to an AI system. When a student asks for practice, *Ethel* can generate small, in-chat tasks that can be solved interactively: small multiple-choice or numerical questions rooted in the course material.

7. Selective Automation in Grading: What We Learned at Scale

Grading deep, authentic assessment tasks is resource intensive, usually requiring large numbers of teaching assistants and days of work for large-enrollment courses. Instructors grudgingly resort to solution-oriented formats like multiple-choice or short-answer. Here, AI can assist to 'go back' to paper-and-pencil exercises, as it can assist in grading.[13] We have run experiments in large-enrollment thermodynamics, calculus, and general chemistry exams, evaluating open-ended handwritten work on rubrics. After confidence estimates based on test theory, we were able to assert AI-based point assignments for considerable portions (47–85%) of the grading load with an accuracy comparable to human grading.[14] Fig. 2 illustrates this for a general chemistry exam.[15]

The goal is to take the plunge from experiments to production usage in an upcoming semester while expanding the use of authentic assessment. This reinforces the message that process

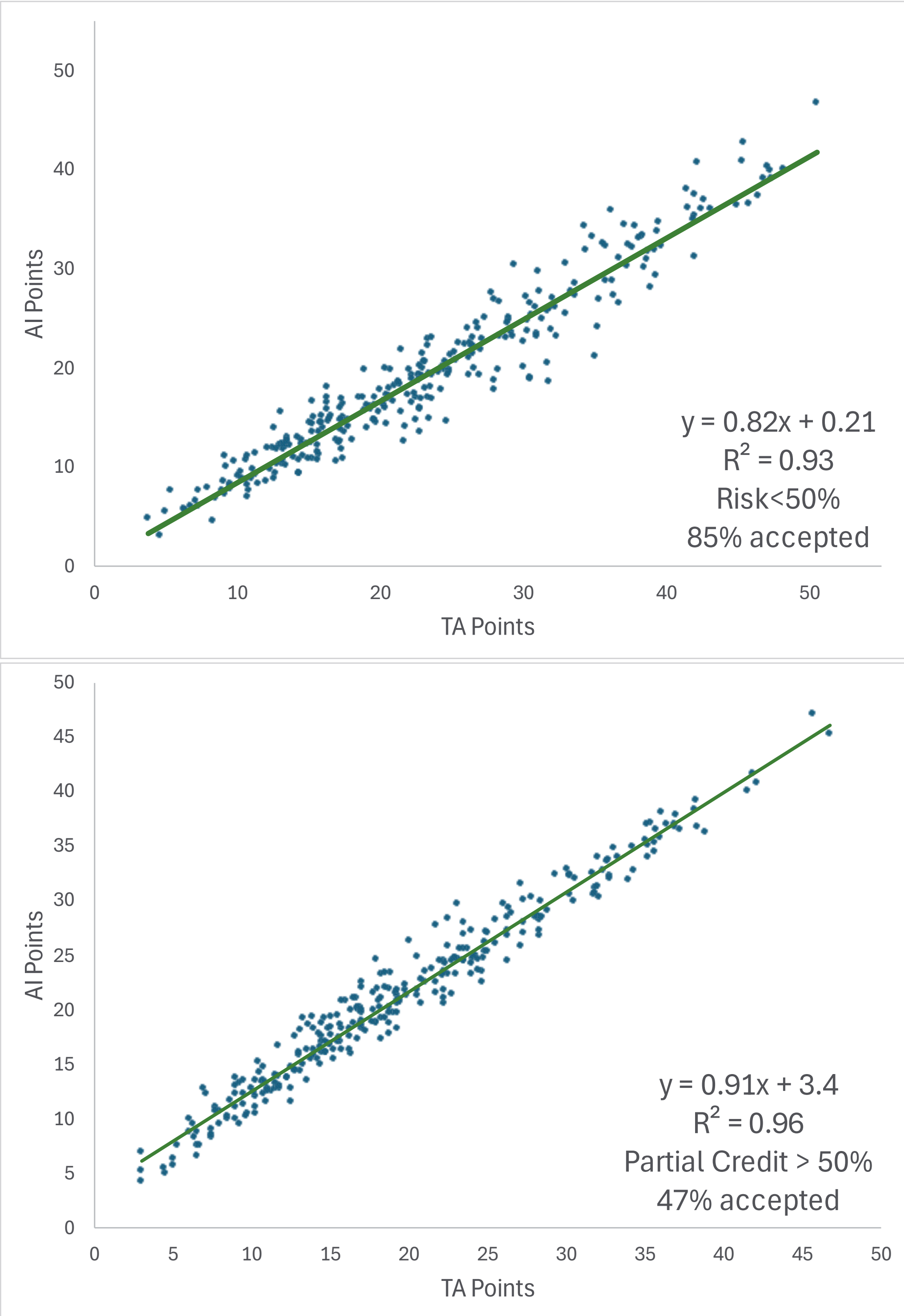

Fig. 2. Relationship of AI-assigned and TA-assigned points for a handwritten general chemistry exam, using two confidence measures. Using the discrepancy between AI-assigned grade and test-theoretical expectation as a measure of risk (left panel), 85% of the grading load could be absorbed by AI; using the simpler measure of partial correctness (right panel), higher accuracy can be achieved at the cost of having higher human workload (47% AI versus 53% human grading).

moves (plans, checks, uncertainty) actually count. If students do not agree with the AI decisions, the plan is to simply let them veto the judgment on particular rubric items, in which case a human will make the decision. Our experiments consistently showed that the system has a very low rate of false positives, so we do not risk artificial grade inflation through artificial intelligence.

8. Transparency About AI Use as a Learning Target

Hidden AI use corrodes both fairness and metacognition. There is a danger that prohibiting AI-usage ends up punishing the honest students, since tools that are aimed at detecting AI-usage are largely flawed,[16] and we should not set rules that cannot be enforced. Besides, in several situations it is as inauthentic to prohibit AI-usage as it is to prohibit using a calculator to determine the square root of 546253 – who does that? In the end, of course, the students are fully responsible for their work, so we should normalize use and grade the quality of verification.

Depending on instructor preferences and context, major assignments might require a short appendix with representative prompts and outputs, a critique of what was correct or spurious, and how claims were verified against course materials or calculations. And before you ask: of course, AI-assistance was used in writing this essay.

Because *Ethel's* answers are grounded in the course corpus with citations, students can learn a disciplined habit: use a tool → inspect evidence → verify → revise. In practice, this policy reduces the cat-and-mouse of 'detection' and turns AI into an object of analysis rather than a shortcut.

9. Supervised and Oral Components for Certification

Short, structured oral checks and whiteboarding remain among the most efficient ways to see how students are thinking. We might want to replace a slice of auto-graded credit with brief, structured oral checks with whiteboarding ('micro-orals') that sample recent material:[17] students sketch a plan, make one representation shift, and talk through a limiting case or an uncertainty source. *Ethel* helps students prepare: simulating *viva*-style prompts drawn from the course notes (see Fig. 3 for an example), while the certification itself is still human-to-human.

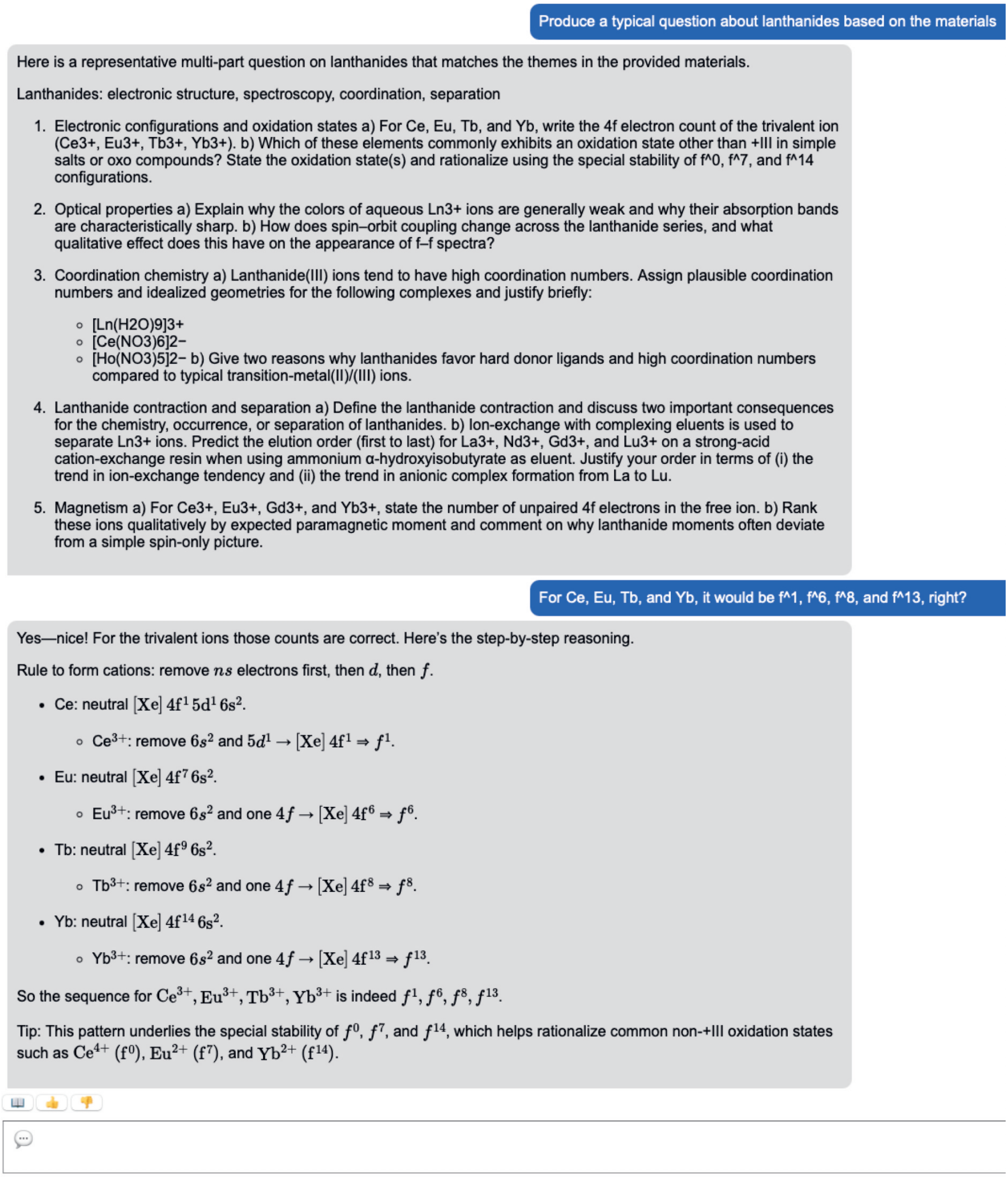

Fig. 3. *Ethel* generates questions based on the materials from an inorganic chemistry course.

10. Learning Together Again: Belonging, Wellbeing, and Durable Habits

The current cohort carries the aftershocks of pandemic isolation, the social pressure and echo-chamber dynamics of online life, and a new ambient message: 'the AI can already do this better than you.' None of that is a recipe for curiosity or persistence. A robust literature links social belonging, peer explanation, and structured group work to learning gains and to students' willingness to stay with hard problems.[18] Peer instruction, studio-style classes, and cooperative problem-solving help students practice argumentation under challenge; they also counter loneliness and 'giving up' by building a community where questions are normal and effort is visible.

An implementation might design weekly rhythms that make the human part non-negotiable. In class, students might work in small groups on critique-and-repair tasks or mechanism justifications; between classes, they use the chatbot to check understanding with citations, not to harvest answers. We might also schedule 'internet-off' intervals: windows where students must produce plans and checks unaided, followed by a compare-against-AI reflection. The message is that tools can help, but growth comes from wrestling with ideas in the company of other people. That stance also supports mental health: clear expectations, visible progress on valued skills, and regular peer contact reduce the helplessness that fuels burnout.

11. Pulling it Together

Across these moves, the through-line is simple: we realign assessment with the understanding we value: process visibility, representation agility, justification, and uncertainty management, because those are the levers that research says improve learning. AI enters as infrastructure: a course-grounded assistant that scaffolds planning, supplies targeted practice, speeds feedback on the right kinds of items, and documents its own role so students can learn to verify rather than outsource. The result is an environment where grading remains fair, learning remains authentic, and students remain connected to the material, to each other, and to the purpose of doing science in the first place. Together, these maintain the heat (authentic tasks, frequent feedback, scalable workflows) while preventing the 'bump' (over-reliance on brittle formats, false certainty, and opaque automation).

12. Outlook

At ETH Zurich, we will continue to develop the *Ethel* ecosystem. In our work, it became apparent that the impact of AI on education goes far beyond chatbots; very often, more advanced uses include multi-stage workflows that include branching points and non-AI stages. Orchestrating these flows and deploying them at scale demands extensive systems engineering, much of it has little or nothing to do with AI; instead, modern-day Large Language Models (LLMs) can be treated as a commodity to provide 'intelligence' at different points of a larger task.

13. Conclusion: Keep the Flame On, Just Keep Swirling

AI has raised the temperature on the familiar tasks of first-year science; our response is not to turn down the heat but to swirl with purpose. If we grade the how (plans, representations, checks, uncertainty) alongside the what, shift weight towards authentic tasks that demand justification and critique, and use automation selectively – with confidence gating, audit trails, and transparent student use – then AI becomes infrastructure for better teaching rather than a shortcut around it. Our exam pilots show that this is feasible at scale: turnaround improves, human attention moves to the items that need judgment, and students see that process actually counts. Just as important, we rebuild the social conditions for learning – peer instruction, studio work, micro-orals, and regular human contact – to counter loneliness, fatalism, and the sense that 'the AI is better than me.' The point is not to outscore the model; it is to cultivate the habits of mind and the communities that make science worth learning. *Keep the flame on; keep swirling.*

Acknowledgements

The author would like to thank Jan Cvengros for making available his exam, productive discussions, and grading the AI-completed exams.

Received: September 9, 2025

[1] G. Kortemeyer, *Phys. Rev. Phys. Educ. Res*. **2023**, *19*, 010132, https://doi.org/10.1103/PhysRevPhysEducRes.19.010132.
[2] G. Kortemeyer, M. Babayeva, G. Polverini, R. Widenhorn, B. Gregorcic, *Phys. Rev. Phys. Educ. Res*. **2025**, *21*, 020101, https://doi.org/10.1103/98hg-rkrf.
[3] R. Kahaleh, V. Lopez, *Phys. Educ*. **2025**, *60*, 025013, https://doi.org/10.1088/1361-6552/adb235.
[4] G. Polverini, J. Melin, E. Önerud, B. Gregorcic, *Phys. Rev. Phys. Educ. Res*. **2025**, *21*, 010154, https://doi.org/10.1103/PhysRevPhysEducRes.21.010154.
[5] B. W. Lloyd, J. N. Spencer, *J. Chem. Educ*. **1994**, *71*, 206, https://doi.org/10.1021/ed071p206.
[6] M. Cooper, *J. Chem. Educ*. **2010**, *87*, 231, https://doi.org/10.1021/ed800096m.
[7] I. Parchmann, S. Hansen, P. Diemel, T. Heyn, A. Grubert, A. Vollersen, *CHIMIA* **2023**, *77*, 668, https://doi.org/10.2533/chimia.2023.668.
[8] M. Cooper, M. Klymkowsky, *J. Chem. Educ*. **2013**, *90*, 1116, https://doi.org/10.1021/ed300456y.
[9] G. Kortemeyer, W. Bauer, *Phys. Rev. Phys. Educ. Res*. **2024**, *20*, 010145, https://doi.org/10.1103/PhysRevPhysEducRes.20.010145.
[10] G. Kortemeyer, *Phys. Teach*. **2024**, *62*, 698, https://doi.org/10.1119/5.0231187.
[11] M. Cooper, R. Stowe, *Chem. Rev*. **2018**, *118*, 6053, https://doi.org/10.1021/acs.chemrev.8b00020.
[12] G. T. Prins, A. M. W. Bulte, J. H. van Driel, A. Pilot, *Int. J. Sci. Educ*. **2008**, *30*, 1867, https://doi.org/10.1080/09500690701581823.
[13] G. Kortemeyer, J. Nöhl, D. Onishchuck, *Phys. Rev. Phys. Educ. Res*. **2024**, *20*, 020144, https://doi.org/10.1103/PhysRevPhysEducRes.20.020144.
[14] G. Kortemeyer, J. Nöhl, *Phys. Rev. Phys. Educ. Res*. **2025**, *21*, 010136, https://doi.org/10.1103/PhysRevPhysEducRes.21.010136.
[15] J. Cvengros, G. Kortemeyer, *arXIV* **2025**, 2509.10591v1, https://doi.org/10.48550/arXiv.2509.10591.
[16] C. Chaka, *J. Appl. Learn. Teach*. **2024**, *7*, 115, https://doi.org/10.37074/jalt.2024.7.1.14.
[17] D. Ramella, *ACS Symp. Series* **2019**, *1340*, 79, https://doi.org/10.1021/bk-2019-1340.ch006.
[18] A. Dao, S. Pegg, S. Okland, H. Green, A. Kujawa, *J. Am. Coll. Health* **2024**, https://doi.org/10.1080/07448481.2024.2355162.