

# SPG Mitteilungen

## Communications de la SSP

**Auszug - Extrait**

### **Dealing with Messiness in Introductory Physics Teaching: A Field of Application of Large Language Models**

*Gerd Kortemeyer, Rectorate and AI Center, ETH Zürich*

This article has been downloaded from:  
[https://www.sps.ch/articles/various\\_articles/](https://www.sps.ch/articles/various_articles/)

DOI: [10.5281/zenodo.13208971](https://doi.org/10.5281/zenodo.13208971)

# Dealing with Messiness in Introductory Physics Teaching: A Field of Application of Large Language Models

Gerd Kortemeyer, Rectorate and AI Center, ETH Zürich

## 1. Introduction

Artificial Intelligence (AI) has arguably been around since 1956 [1], and chatbots have been around since 1965; some of us, the author included, may have even programmed versions of ELIZA [2] on their original IBM PCs (the one with the click-clack keyboard) or their Apple IIs. Other systems followed, some of them equally deterministic (essentially a collection of if-statements), some of them more probabilistic, relying on statistical methods to make predictions and decisions. The latter incorporate uncertainty and randomness, and they are oftentimes better suited to dealing with complex, real-world scenarios where data is incomplete, noisy and messy; Large Language Models (LLMs) like GPT fall into this category [3].

As physicists, we are generally far from computer-phobic, and we incorporate a wide spectrum of computing into all aspects of our work, except for two concepts we may have been somewhat snobbishly avoiding: SQL-databases (“too bookkeepish”) and AI (“too messy”). AI in general, and even Machine Learning (ML) in particular, may have been dismissed as only useful if one has no clue what is really going on. With the notable exceptions of having to deal with the vast amounts of data produced by for example particle physics or astronomy, by its very nature, ML is mostly phenomenological and thus dangerously close to messy fields like biology (because life is so messy). Physics, being highly theoretical and mathematical, has its own methods of understanding the universe that might not always align with AI’s data-driven approach. And in any case, when it gets too messy, it is time to resort to the spherical cow [4].

So, why would we suddenly want to deal with messy AI when it comes to education? I would like to propose the following argumentation chain: the active pursuit of learning is maybe one of the most human activities; humans are messy (and neither spherical nor bovine); thus, when it comes to deploying educational technology in pursuit of helping humans learn, it needs to be the messy kind of technology – like LLMs. In the following, this idea of essential messiness, and of the potential role of AI related to it, is discussed for four aspects of undergraduate physics teaching: language, problem solving, programming, and cheating.

## 2. Chatbots and Messy Language

Physics uses very precise language, and communicating physics relies on adhering to definitions and terminology; different from everyday language, words like momentum, energy, and power are not synonymous. We also use arcane expressions like “fall off quadratically” or “to first order;” it is a learning goal of introductory physics to teach this language, but students will not be there yet [5,6]; while still exploring, students need to be able to speak like the novices that they are [7]. In a survey at ETH Zurich, many students stated that they use LLMs to get an entry into a new field if they do not even know how to ask the right questions [8]; they appreciated that they do not need to be embarrassed to ask ChatGPT “dumb questions.” LLMs can deal with the messiness.

Students are aware that all LLMs hallucinate and are generally willing to put up with that; however, they also complain that a two-class society is developing between students who can afford the subscription-based versions and the much less powerful openly available models. They thus called for

ETH to provide a more reliable model for its courses, which would be available for everybody [8]. Consequently, ETH started an experiment called Ethel, which would offer custom-chatbots for courses based on actual course materials: scripts, exercises, slides, lecture recordings, etc. The bots use a method called Retrieval Augmented Generation (RAG) [9], where reliable documents are made available as reference while LLMs deal with the

1. Die Ausbreitungsgeschwindigkeit nimmt mit der  $\lambda$  ist, desto schneller kehren die Massenelemente in
2. Die Ausbreitungsgeschwindigkeit nimmt mit der  $\lambda$  ist, desto langsamer kehren die Massenelemente in

### 1.2.9 Energiedichte

Wir wollen uns nun der **Energiedichte**  $dW/dV$  von men. Diese setzt sich aus der **kinetischen Energiedichte**  $dE_{kin}/dV$  zusammen, welche die poten der einzelnen Massenelemente des Mediums untereinander Mit den Ergebnissen aus dem vorherigen Abschnitt kö dichte einer transversalen Seilwelle  $\xi(x, t)$  herleiten. Aus dem Physik I Kurs ist bekannt, das die kinetische der Welle gegeben ist durch

$$dT = \frac{1}{2} v^2 dm,$$

wobei  $v = \frac{\partial \xi(x,t)}{\partial t}$  die Geschwindigkeit des Massenelemente Mit der Beziehung für die Massendichte  $\rho = \frac{dm}{dV}$  folgt f

$$\frac{dT}{dV} = \frac{1}{2} \rho \left( \frac{\partial \xi}{\partial t} \right)^2$$

Wir werden nun die elastische Energiedichte  $dE_{el}/dV$  spannung  $S = F/A$  wirkt, ist die in einem um die Läng  $\xi$  *conservierte potentielle Energie* gegeben durch  $dE_{el} =$  **Retrieval Augmented Generation in action: the right window shows a dialogue with Ethel, which is based on the information in the lecture script on the left.**

Explain that example about waves on a rope, I don't get what that "dm" is about.

The term "dm" in the context of the wave on a rope example represents a small mass element of the rope. When waves propagate in a medium like a rope, the energy of the wave is distributed among small pieces or elements of that medium.

In the given example, they're considering a small segment of the rope with mass "dm". This segment is small enough that it can be treated as a point mass for the purpose of calculations, but large enough that it contains a very large number of atoms so macroscopic quantities like mass can still be defined.

The kinetic energy of this mass element "dm" is given by the equation

$$dT = \frac{1}{2} v^2 dm,$$

where  $v = \frac{\partial \xi(x,t)}{\partial t}$  is the speed of the mass element, and  $\xi(x, t)$  is the displacement of the rope at position  $x$  and time  $t$ .

The kinetic energy density (kinetic energy per unit volume) is then obtained by dividing this kinetic energy by the volume of the mass element, which is  $dV$ . With  $\rho = \frac{dm}{dV}$  as the mass density of the rope, the kinetic energy density becomes

$$\frac{dT}{dV} = \frac{1}{2} \rho \left( \frac{\partial \xi}{\partial t} \right)^2.$$

I hope this clarifies the role of "dm" in the energy calculations for a wave on a rope. Let me know if you have any further questions!

The "v" in your answer seems different from the "v" we got out of the partial differential equation for  $\xi(x,t)$  - is that the velocity of the wave?

Yes, the "v" in the context of the equation for  $\xi(x, t)$  is indeed the velocity of the wave, more specifically, the phase velocity. This is the speed at which a particular phase of the wave (for instance, the crest) propagates in space. It can be represented by the equation  $v = \frac{\omega}{k}$ , where  $\omega$  is the angular frequency of the wave and  $k$  is the wave number.

In the kinetic energy equation,  $v$  represents the speed of the small mass element "dm" of the rope, which is a function of time and position on the rope and is related to the displacement of the rope.

These two "v"s, while represented by the same letter, refer to different concepts. The context should make it clear which one is referred to.

conversational aspects (and bridge language-differences, as they can deal with a wide spectrum of spoken languages).

The figure shows a dialogue with Ethel. The lecture script has over 300 pages and is written in German; in addition, LaTeX is understood and spoken by the bot. RAG allows learners to ask specific questions about the lecture materials in multiple languages, and it does not require exact terminology or search phrases (i.e. allows for “messiness”); as everything is in course context, students can simply refer to “that example about waves on a rope” and resolve confusion about terms or symbols in the script (e.g., “the ‘v’ we got out of the partial differential equation for  $\eta(x,t)$ ,” where “eta” is not even the correct Greek character). The LLM acts as a conversational and reasoning agent to deal with the messiness, while the knowledge comes from the materials, leading to less hallucinations.

The setup uses on-campus servers in connection with GPT-4 running on Swiss datacenters under the same licensing terms used for other cloud services dealing with confidential data [10]. This setup allows for faster dialogues than GPTs [11], and it has no hard limit on how much reference data is used. We received very encouraging feedback to this simple approach from both students and faculty, and we plan on further developing it, e.g., course materials getting automatically “digested” (embedded) as soon as faculty adds them. Future versions might use a fine-tuned, open-weight, and potentially on-premises models like Llama, Mistral, or Gemma instead of a commercial backend. ETH Zurich and EPF Lausanne are currently collecting “donations” of course materials from faculty for tuning; it is an experiment, and it remains to be seen if this works.

### 3. Problem-Solving and Messy Derivations

While most of us physicist likely enjoyed physics lectures and appreciated the inspiration, fun, role-modeling, and valuable perspectives gained (which we aim to pass on to our students), we would probably also admit that we learned more physics facts and methods from solving problems – on paper and in the lab. The process of deriving and arriving at a solution, including all the detours and errors, is what in the end helped us master the concepts; the journey is the goal. Educational technology can help provide more opportunities for learning by assessment, also known as formative assessment; unfortunately, many of these technologies only focus on the final answer (“42 Newton”) instead of the way how to get there, which for a learner is the most important part. The reason: “42 Newton” is nice and clean, and one can check this final result deterministically, but the non-linear journey towards it might be a mess.

While the mathematical capabilities of LLMs still leave a lot to be desired (for example, do not count on them being able to count), the text corpus used for the training of the most popular ones is extensive enough in the foundational sciences to derive the solutions of about 80 % of introductory physics problems, including freshly authored ones that are guaranteed not to be part of the training [12,13]. Rather than seeing the challenge of students cheating on homework, we should see the opportunity of providing immedi-

ate, individualized feedback or even assisting in grading student-provided derivations.

In a feasibility study, we tested AI-grading of a synthetic dataset of homework solutions, looking at the complete answer derivation, and found an  $F^2$  of 0.84 between AI-grades and instructor grades on a grading rubric considering correctness of argument, straightforwardness, symbolic correctness, and numerical correctness of the solution path [13]. While not good enough for high-stakes summative exams, the result encouraged us to provide automated feedback on handwritten formative assessment (“exercise sheets”) in an introductory physics course. As students are handing in their work as scanned PDFs anyway for teaching-assistant feedback, we used MathPix [14] to transcribe them into machine-readable format and GPT-4 for AI-feedback on the solution, based on the problem and the instructor-provided sample solution. The feedback was not always correct (which students are aware of [8]), and we had to work on the “character” of this virtual teaching assistant by refining the prompts, but in the end, we received nothing but encouraging evaluations from the students.

Currently we are working with a high-stakes thermodynamics exam, where we aim to put the complete workflow to the test: scanning of exam sheets, transcription with MathPix and GPT-4V, and rubric grading. This data set of 252 exams with four multi-part problems with a detailed step-by-step grading rubric is hopefully going to give us a better understanding of uncertainty and confidence levels.

A major caveat: as with so many of the popular cloud services, there are legal challenges for their production usage. While GPT-4 and GPT-4V are covered by existing agreements of ETH Zurich, we would still need to attend to the contractual details with MathPix, should we decide to use this service outside the research protocol and without informed consent by the learners.

### 4. Messiness and Programming

Over the recent decades, Computational Physics have found an entry into introductory physics courses [15]. While visualizing and “playing with” physics using simulations and computerized data collection is widely acknowledged as beneficial [16], a common concern of educators is that struggling with the syntax of for example VPython could easily distract from the physics; the author regularly used VPython in his courses and can attest to the potential of a parenthesis being in the wrong place to take up a disproportionate amount of instructional time. LLMs can deal with messy instructions, assist in debugging, and produce syntactically correct code. At the same time, the LLM might make some of the very same physics errors that students do (like, “forgetting” to norm a unit vector [12]); thus, LLMs can overcome hurdles imposed by syntax (which are arguably not part of the learning objectives of an introductory physics course) in favor of concentrating on and debugging the physics.

### 5. The Messiness of Cheating

With LLMs performing at the level of a B-student in introductory physics, a large concern is cheating; will students still

do the exercises themselves, or will they have GPT generate their answers, as it is frequently bemoaned in the humanities with its take-home essays?

According to the student survey at ETH Zurich, students make extensive use of AI-tools for a variety of tasks, many of them language-oriented, such as grammar checks, translations, summaries, and feedback on writing. It is also frequently used for first drafts of computer programs, debugging, and explaining code. For other tasks in the STEM disciplines, GPT was deemed too unreliable, and in any case, students remarked that they were worried about not learning the materials [8]. In this context, it is important to note that course grades and study progress are mostly determined by high-stakes exams, which take place under controlled conditions on-campus, often with just paper-and-pencil and a non-programmable calculator. It is thus not worth it to cheat on homework exercises during the semester.

The situation is different in university systems where grades are determined by a variety of small assignments over the course of the semester. Here, cheating is and likely always was rampant, using commercial non-AI sites like Chegg [17] or simply copying from peers. There is evidence that as of the end of 2023, student behavior in physics has hardly changed with the availability of AI-tools, and traditional means are still strongly preferred [18].

On the other hand, tools like Chegg have a delay till solutions become available, which to a certain extent makes them “exam safe” if the problems had not previously been published. While having remote exams in physics might have been irresponsible before, this practice, which established itself during COVID-19, is certainly irresponsible now [18]; short of total surveillance via cameras and microphones, no number of other lockdown tools will help protect the honest students (which are in the majority).

The findings on cheating are distressing, since there is substantial evidence that repeated exposure to and interaction with the materials, including frequent in-semester assessment, rather than cramming during exam preparation, yields better learning outcomes [19-21]. LLM feedback from highly scalable and quick handwritten in-class quizzes might be key to resolving this dilemma; besides, personalized feedback from these systems might be attractive enough to make cheating appear undesirable and unprofitable.

## 6. Conclusion

As physicists, we may have been spared from having to fundamentally rethink our curricula because LLMs are so bad in math. This flaw is intrinsic, since these systems are calculators for words, not numbers. However, LLMs can be trained to use helper applications; GPT-4 already makes extensive use of Python and R to figure out the math that it cannot do itself (when “analyzing” shows up in a dialogue, tools like sympy are being invoked and arrive at the correct solution, which then gets embedded into the response). Sooner or later, we will be in the same shoes as Computer Science and the humanities, and we have to rethink where

we are going with introductory physics courses. In that respect, LLMs are like pocket calculators: students used to spend huge amounts of instructional time learning how to do things like square roots on paper – nobody does that anymore except for numbers like 49 or 121 and their immediate vicinity, since we already carry what used to be supercomputers in our pockets or on our wrists. Our introductory physics curricula, essentially unchanged since the modern physics revolution, just got a little messier.

## References

- [1] Allen Newell and Herbert A. Simon, The Logic Theory Machine: A Complex Information Processing System, Rand Corporation Report No. **P-868** (1956)
- [2] Joseph Weizenbaum, ELIZA - a computer program for the study of natural language communication between man and machine, *Communications of the ACM* **9** (1), 36-45 (1966)
- [3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin, Attention is all you need, *Advances in neural information processing systems* **30** (2017)
- [4] Margaret Harris, Consider a Spherical Cow, *Physics World* **25**(11), 68 (2012)
- [5] Edward F. Redish and Eric Kuo, Language of physics, language of math: Disciplinary culture and dynamic epistemology, *Science & Education* **24**, 561-590 (2015)
- [6] Paul W. Irving, Daryl McPadden, and Marcos D. Caballero, Communities of practice as a curriculum design theory in an introductory physics class for engineers, *Physical Review Physics Education Research* **16**(2), 020143 (2020)
- [7] Martin Wagenschein, Die Sprache im Physikunterricht, *Zeitschrift für Pädagogik* **7** (Beiheft), 125-142 (1968)
- [8] Fadoua Balabdaoui, Nora Dittmann-Domenichini, Henry Grosse, Claudia Schlienger, Gerd Kortemeyer, A survey on students' use of AI at a technical university, *Discover Education* (submitted)
- [9] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, *Advances in Neural Information Processing Systems* **33**, 9459-9474 (2020)
- [10] Gerd Kortemeyer, Embeddings and RAG with Azure OpenAI API, ETH Cloud Services Blog, <https://cscblog.ethz.ch/index.php/2024/02/06/az-open-ai-rag-chromadb-langchain/> (accessed March 24, 2024)
- [11] OpenAI, Introducing GPTs, <https://openai.com/blog/introducing-gpts> (accessed March 24, 2024)
- [12] Gerd Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, *Physical Review Physics Education Research* **19**(1), 010132 (2023)
- [13] Gerd Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study, *Physical Review Physics Education Research* **19**(2), 020163 (2023)
- [14] MathPix, <https://mathpix.com> (accessed March 2024)
- [15] Ruth Chabay and Bruce Sherwood. Computational physics in the introductory calculus-based course, *American Journal of Physics* **76**(4), 307-313 (2008)
- [16] Edward F. Redish, The computer revolution in physics education? It's here!, *American Journal of Physics* **91**(6), 415-415 (2023)
- [17] Chegg, <https://www.chegg.com/> (accessed April 2024)
- [18] Gerd Kortemeyer and Wolfgang Bauer, Cheat sites and artificial intelligence usage in online introductory physics courses: what is the extent and what effect does it have on assessments?, *Physical Review Physics Education Research* (submitted)
- [19] James T. Laverty, Wolfgang Bauer, Gerd Kortemeyer, and Gary Westfall, Want to reduce guessing and cheating while making students happier? Give more exams!, *The Physics Teacher* **50**(9), 540-543 (2012)
- [20] Daniel T. Seaton, Gerd Kortemeyer, Yoav Bergner, Saif Rayyan, and David E. Pritchard, Analyzing the impact of course structure on electronic textbook use in blended introductory physics courses, *American Journal of Physics* **82**(12), 1186-1197 (2014)
- [21] Louis Deslauriers and Carl Wieman, Learning and retention of quantum concepts with different teaching methods, *Physical Review Special Topics - Physics Education Research* **7**(1), 010101 (2011)