Letter from the Newsletter Editor

Dear GDS community,

I hope you enjoy this newsletter, which was curated to not only inform you of what GDS has been up to but also to learn more about the world of data science and physics. The articles this time are focused on Large Language Models (LLMs) that will hopefully be food for thought.

There are also announcements regarding GDS activities including the APS Fellows and the March and April meeting sessions.

Please share the newsletter and consider submitting ideas or better, articles, in the future!

Best wishes,

Alexis V. Knaub
Editor
Message from the Chair

Dear GDS members,

I want to start by thanking you for the opportunity to serve as your Chair!

In a recent discussion, Professor Julie Butler, our Social Media officer, presented a graph showing the number of papers on the arxiv associated with physics and machine learning as a function of time.

![Graph showing the number of papers on the arxiv associated with physics and machine learning as a function of time.](image)

From this, we can see the explosive growth in papers in our field! In this year, we have seen the growth in the public’s awareness of large language models along with exciting work in fields ranging from particle physics to condensed matter.

His has been another productive year for GDS! We are looking forward to a number of exciting sessions at the March and April Meetings in 2024! Please join us at our business meetings to suggest sessions for next year! A special thanks to our incoming chair, Jerome Delhommelle and our Member-at-Large, Stefano Roberto Soleti for organizing our programs for 2024. I would also like to extend a thanks to all of you who are organizing sessions, sorting abstracts, and contributing to the success of our programs at the March and April meetings! We will also be running a full day tutorial at the March meeting on Data Science in partnership with a number of other units!

We successfully advanced two APS fellows who will be honored at our business meetings at the March and April Meetings! We will also continue with our GDS IMPACT awards for students who have made an impact on the field.

Our efforts to incorporate Data Science into the undergraduate physics curriculum through the Data Science Education Community of Practice (DSECOP) have continued, with workshops and more units developed by our fellows. Thanks to the efforts of our past Chair, Marilena Longobardi, we maintain a strong connection with FIAP and our industrial members. She and our members-at-large Jennifer Hobbs and Nima Dehmamy have a working group on this for any who would like to join.

Finally, I would like to thank the members of the executive committee for their efforts to make GDS a success. Next year, you will be in good hands as Jerome Delhommelle takes over as Chair.

Please encourage your friends and colleagues to join GDS! For a limited time, we are offering free membership to existing members with the code GDS23-FirstYearFree. We are very close to the numbers required to become a division!

Best,

William Ratcliff
GDS Chair
Large Language Models (LLMs) these days can write essays, summarize research papers, generate recipes and travel itineraries, and debug your code — but ask ChatGPT to multiply two four-digit numbers, and it will fail over 90% of the time. [1]

It turns out that numbers are quite distinct from other forms of language. Numbers have specific meanings, but unlike words, these meanings exist on a continuous scale with infinitely many values, all guided by a complex and precise system of rules. Language and mathematics, then, arguably shouldn’t mix, at least within AI models. We clearly shouldn’t expect LLMs to be perfect calculators. We can even employ clever strategies to help mitigate this issue: we can give LLMs access to calculators and code interpreters, for instance, or we can guide LLMs to break down challenging mathematical questions into simpler ones.

That said, I can’t help but think that there is something thrilling about this apparent incompatibility between language and mathematics in the current landscape of AI. In a way, my role as a scientist is to translate data into language, albeit with a toolkit based on mathematics and code. Similarly, I think there are compelling reasons why we might want to tackle the challenge of blending language and data in our scientific AI models as we envision how the act of doing science could change in the next 5-10 years.

For instance, how might science change if scientists had access to an AI model trained on a massive variety of scientific data? LLMs achieve a fluency with language-based tasks, even ones they weren’t explicitly trained on, because they have been trained on over an astounding amount of text data from diverse sources on the Internet. Would an AI model of such scale specializing in numerical data open similarly innovative paths of inquiry for scientists in the near future?

One key reason why we haven’t yet seen major models like this emerge is that scientific datasets come in highly specialized formats that require domain expertise to understand. Most of the so-called “foundation models” we see shaping the public’s experience of AI today are experts in a single data format: text, images, video, etc. Similarly, AI models in science today are carefully constructed to reflect the highly-curated datasets on which they are trained. A model spanning scientific domains, however, needs to be very flexible — as flexible as an LLM, yet grounded in a rigorous sense of numerics.

Every proposal for how to treat numbers in language models struggles with how to translate the infinite space of numbers into a finite number of vocabulary elements. LLMs break down language into pieces called “tokens”, sort of like tiles in a game of Scrabble. Introducing numbers into this framework can be like adding an infinite number of Scrabble tiles into the mix, making the game impossible to play.

Existing strategies that map numbers to language tokens struggle to generalize beyond tokens they’ve already seen in the training data. A strategy called xVal [2] from my colleagues and I at Polymathic AI [3] proposes a way to make LLMs end-to-end continuous when processing numerical values, there-
by improving their ability to interpolate between known values. We suggest that small improvements like this could help adapt language models to one day be more suitable for training on diverse scientific datasets.

LLMs have opened up creative ways of reading, writing, and learning by responding to text-based prompts in a much more tailored way than what we’ve seen before. In their current form, they are clearly incompatible with scientific data: they don’t understand mathematics, fabricate outputs, and can propagate dangerous biases in a training dataset. But language models also provide a rare flexibility of input format that could be key to building scientific foundation models that connect truly diverse domains. We shouldn’t discount this connection between language and mathematics when imagining the next generation of scientific tools: after all, language is ultimately what gives our discoveries meaning.

3. https://polymathic-ai.org
Natural language processing (NLP) is a field as old as computer science; its origins originally centering around the post-WWII era need for translations from Russian to English. With the advent of neural networks, the Word2Vec representation, and the internet era of big data providing a ready made Rosetta Stone, the early 2010’s saw a dramatic surge in NLP performance. The invention of transformer architectures, a doubling down on model size in the form of hundreds of billions of parameters, and human feedback-based reinforcement learning have pushed the term Large Language Models (LLMs) into the mainstream lexicon.

Within education the rise of LLMs has caused a great deal of hand wringing and blanket bans owing to the unexpectedly superior performance of GPT-4 on a number of standardized college level exams. There have even been, since disproven, claims that LLMs could ace entire college curricula if properly prompted. At the heart of the issue, is a fear of rampant unidentifiable academic fraud, which would jeopardize the assessment of learning outcomes and create a LLM-reliant workforce.

Rather than sticking our heads into the sand, we should understand that LLMs are human-enhancing tools that have been broadly available for less than a year, will not go away, and can be used to increase students’ level of abstract reasoning. Through conversations with my colleagues and students, I have arrived at the following (partial) list of ways that I currently use LLMs in an educational setting.

### LLMs As a Classroom Enhancer:
Coding is foundational to 3rd and 4th paradigm education and has traditionally been a key source of inequity between students from disadvantaged backgrounds. LLMs break down this barrier by allowing the students to explain their desired functionality in natural language and then iteratively generate functional code in most common programming languages. This removes a barrier to entry for the students and permits the professor to place more focus on algorithmic thinking in class and on assignments.

Similarly, instructors should embrace the tendency of LLMs to generate confident sounding responses by encouraging their students to use LLMs to answer homework or exam questions. Quite often the LLMs are incorrect and the types of errors can span the range from simple calculation errors to flawed (memorized) reasoning. As professors we understand nothing deepens your understanding of a topic as much as finding the source of an error in a students’ solution; our students would learn more by performing the same exercise of LLM generated responses.

### LLMs As Learning Accelerators
One of the greatest strengths of LLMs is their ability to effectively summarize text and provide those summaries with varying levels of subtlety. Multi-prompt conversations with LLMs can be used to get an overview of a new field, learn keywords for searching, and even to recommend books or technical papers to read. Of course, these recommenda-
tions are peppered with hallucinations but I find that often they make for a good starting point. Even if the articles recommended don't exist the listed authors are often luminaries within the field.

Text summarization does not need to be limited to zero-shot interrogations of the memorized text used to train the LLM. There have already been fine-tuning studies where additional text is provided to prompts for meta-studies of a field and there is promise for large scale meta-summaries of entire fields of the technical literature. The underlying irreproducibility of LLMs means that grounding is important for conclusions drawn from such studies but we already regularly use these tools in our own work.

**LLMs in Academic Writing**

Translating complex physics concepts into text written at the appropriate level can be challenging, sometimes even moving from a bulleted outline to text on a page can be an angsty process. LLMs are masters of taking outlines and converting them seamlessly into narrative. A LLM can be used to explore writing tones, methods of presenting ideas, rephrasing text to address different audiences, and clarifying ambiguous text. My favorite prompt when writing proposals has become “Rephrase the following text for a non-specialist with an advanced degree.” A LLM’s tendency to incorporate vague and flowery text means that their text can’t be used as is, but it often contains interesting ideas that can be used to improve the original text.

LLMs are now a permanent part of the educator toolkit; we should embrace them for their ability to increase human understanding and not relive the calculator in the classroom arguments of the 1980’s. Now is the time to reimagine how we educate, evaluate and elevate our students (and ourselves) in light of the power of these tools, identify how they can be used effectively and ethically, and make certain that the students we graduate are not competing for jobs supplanted by a new technology.
In the last year, large language models (LLM), based on cutting edge machine learning tools and computational power, have taken the world by storm. Chatbots like ChatGPT and its competitors have reached a wide audience. LLM are trained in an unsupervised way on huge amounts of text. The latest incarnations have many billions or even trillions of parameters which are optimized during training. LLM operate on tokens (words or parts of words) using the attention mechanism: soft weights are calculated for each token in parallel (so called transformers), based on the surrounding tokens. In contrast to hard weights, which are fixed after training, they can change over time when the model is invoked, providing much greater flexibility.

Given input prompts of text, LLM can generate the statistically most likely text outputs, and even images. One problem is that small variations in the inputs can result in quite different results, or even generate patently false, but convincingly looking outputs, aptly named “hallucinations”. Besides the ethical implications for responsible use of LLM, uses in science will benefit from explainability of how the models reach their conclusions, and the ability to reliably trace the sources on which they are based. After all, citations and the avoidance of plagiarism are pillars of the research process.

So, what is their relevance for, and use in, the physical sciences? As is often the case, this is a two-way street: LLM can be a potentially useful research or study tool, and physics can shed light on the ways LLM perform their “magic”.

To begin with, LLM tools can generate, edit or summarize scientific texts. They can be a new type of search engine answering scientific questions in a concise and “ready to use” way. This can be useful both for researchers and students to streamline the gathering and processing of information, especially if some guardrails are provided. For example, tools to detect, or watermarks to show, the provenance of text as LLM-generated will be very useful.

On the other hand, because LLM are trained on existing texts, they will hit limitations when trying to generate “new” knowledge, and even more so when attempting to use them for scientific discoveries or paradigm shifts. The restricted ability of LLM to grasp fine nuances and really “understand” scientific text will naturally minimize their use for peer reviews, if we want to maintain the integrity of the scientific process. And finally, the “language” of physics is mathematics, and LLM, specializing on human languages, are, so far, not very good at math.

In the opposite direction, in analogy to statistical thermodynamics, basic properties of complex systems (like pressure and temperature) can be connected to the properties and interactions of the myriads of constituents in the system. For LLM the tokens are the basic “constituents”, and the soft weights computed during training can be thought of as arising from the “forces” between them. This opens interesting avenues to explore the methods developed in physics, e.g. for phenomena like phase transitions, superfluidity or superconductivity, for better understanding basic features of LLM, and ultimately opening the model “black box” to provide interpretability insights.

To conclude, when used with care, LLM can be a powerful new study and research tool in physics, and provide a novel exciting platform for interdisciplinary research.

Disclaimer: this review was written without the use of any LLM tools.
APS Fellows

Boris Kozinsky [2023]
Harvard University / Bosch Research

Citation: For the development of innovative computational and machine learning methods to study microscopic transport and dynamic phenomena, and for their application to the discovery and understanding of technologically relevant materials for energy storage and conversion.

Nominated by: GDS

Tilman Plehn [2023]
Heidelberg University

Citation: For advocating the use of advanced machine learning and data science tools in theoretical work within fundamental physics, as well as for original research in this area.

Nominated by: GDS
**APS Tutorial, March 2024**

**DATA SCIENCE FOR PHYSICISTS I**

**Who Should Attend:** Graduate students, post-docs and other scientists interested in learning how to apply data science to their research. The lectures will provide an introduction to data science and its applications in physics. We assume that participants will have some experience with Python, Numpy, and Matplotlib at the level of a software carpentry course and we will provide a link to learning materials before the tutorial.

**Tutorial Description:** Data Science is playing an ever increasing role in physics. While some departments have offered courses, many of the examples are in the context of social science and other disciplines. In this tutorial, we will introduce data science in the physics context. We will start by introducing Jupyter notebooks and how to explore and visualize data. We will then introduce unsupervised learning techniques including clustering, random forests, etc. We will conclude with an introduction to neural networks and object tracking.

**Topics covered:**
- Data Visualization and Exploratory Data Analysis
- Unsupervised Learning
- Convolutional Neural Networks

**Organizers:** William Ratcliff, Talat Rahman

**Presenters:**
Julie Butler (University of Mount Union)
Karan Shah (Technische Universität Dresden)
William Ratcliff (NIST)

**DATA SCIENCE FOR PHYSICISTS II**

**Who Should Attend:** Graduate students, post-docs and other scientists interested in learning how to apply data science to their research. The lectures will provide an introduction to data science and its applications in physics. We assume that participants will have some experience with Python, Numpy, and Matplotlib at the level of a software carpentry course as well as comfort with the topics covered in the Data Science for Physicists I tutorial.

**Tutorial Description:** Data Science is playing an ever increasing role in physics. While some departments have offered courses, many of the examples are in the context of social science and other disciplines. In the second part of this tutorial we will cover several advanced applications of data science to physics. We will cover physics-informed neural networks (PINNs), which incorporate the physical properties of partial differential equations into the structure and training of neural networks. Examples of PINNs will focus on geophysical and climate-related problems, but PINNs can be applied in other contexts such as Density Functional Theory. Another topic we will cover is image segmentation. For example, how do we automatically detect, segment, and classify objects in a wide variety of images? Examples will be presented from biophysical contexts, but the methods can be applied to many other fields of physics. We will also cover Bayesian Optimization/Active learning, which leads to better sampling of measurements, and can be applied to both theory and experiment. Our final topic will be Sparse Identification of Nonlinear Dynamics (SINDy), which deals with learning equations from experimental data.

**Topics covered:**
- Physics Informed Neural Networks
- Image segmentation
- Bayesian Optimization
- SINDy/Symbolic Regression

**Organizers:** William Ratcliff, Talat Rahman

**Presenters:**
Ching-Yao Lai (Stanford)
Jan Funke (HHMI Janelia Research Campus)
Austin McDannald (NIST)
Chris Amey (Brandeis)
GDS Executive Committee

Chair
William Ratcliff (03/23–03/24)
National Institute of Standards and Technology (NIST)

Chair-Elect
Jerome P Delhommelle (03/23–03/24)
University of Massachusetts Lowell

Vice Chair
Eun-Ah Kim (06/23–03/24)
Cornell University

Past Chair
Maria Longobardi (03/23–03/24)
University of Basel

Treasurer
Savannah J Thais (03/22–03/25)
Columbia University

Secretary
Benjamin Nachman (03/21–03/24)
Lawrence Berkeley National Laboratory

Assigned Council Representative
James Knox Freericks (01/20–12/23)
Georgetown University

Member-at-Large
Jennifer Hobbs (03/22–03/25)
Zurich North America

Member-at-Large
Trevor David Rhone (03/22–03/25)
Rensselaer Polytechnic Institute

Member-at-Large
Casey Berger (06/23–03/26)
Smith College

Member-at-Large
Jihua Chen (06/23–03/26)
Oak Ridge National Laboratory

Early Career Member-at-Large
Stefano Roberto Soleti (03/22–03/24)
Lawrence Berkeley National Laboratory

Student Member
Karan Shah (06/23–03/25)
Helmholtz Zentrum Dresden-Rossendorf