created with Leonardo AI by Kristian Gonzales Barman

Examining the Scientific Understanding of Large Language Models in Particle Physics

Sascha Caron

(Radboud University, Nijmegen and Nikhef, Amsterdam)

Outline

- Some machine learning for large language models
- What is “scientific understanding” ?
- Scientific understanding for AIs
- Why is this relevant for HEP ?

From Transformers to Large Language Models

A **Large Language Model** is trained to predict the next word on vast context of surrounding words.

Transformers Network architecture have revolutionized Machine Learning in the recent years
→Designed to handle "sequences" of data

Introduced 2017 with the paper
"Attention is all you need"
https://arxiv.org/abs/1706.03762

How does this work in 5 min.

Why?

Massive training/scaling to input data
Massive models
Few shot / transfer learning
(pretrained + then fine-tuned)

➔ *New thing:*
➔ *Large Scale pretrained*
*“**Foundation** Models” (Dall-e, BERT, GPT etc.)*

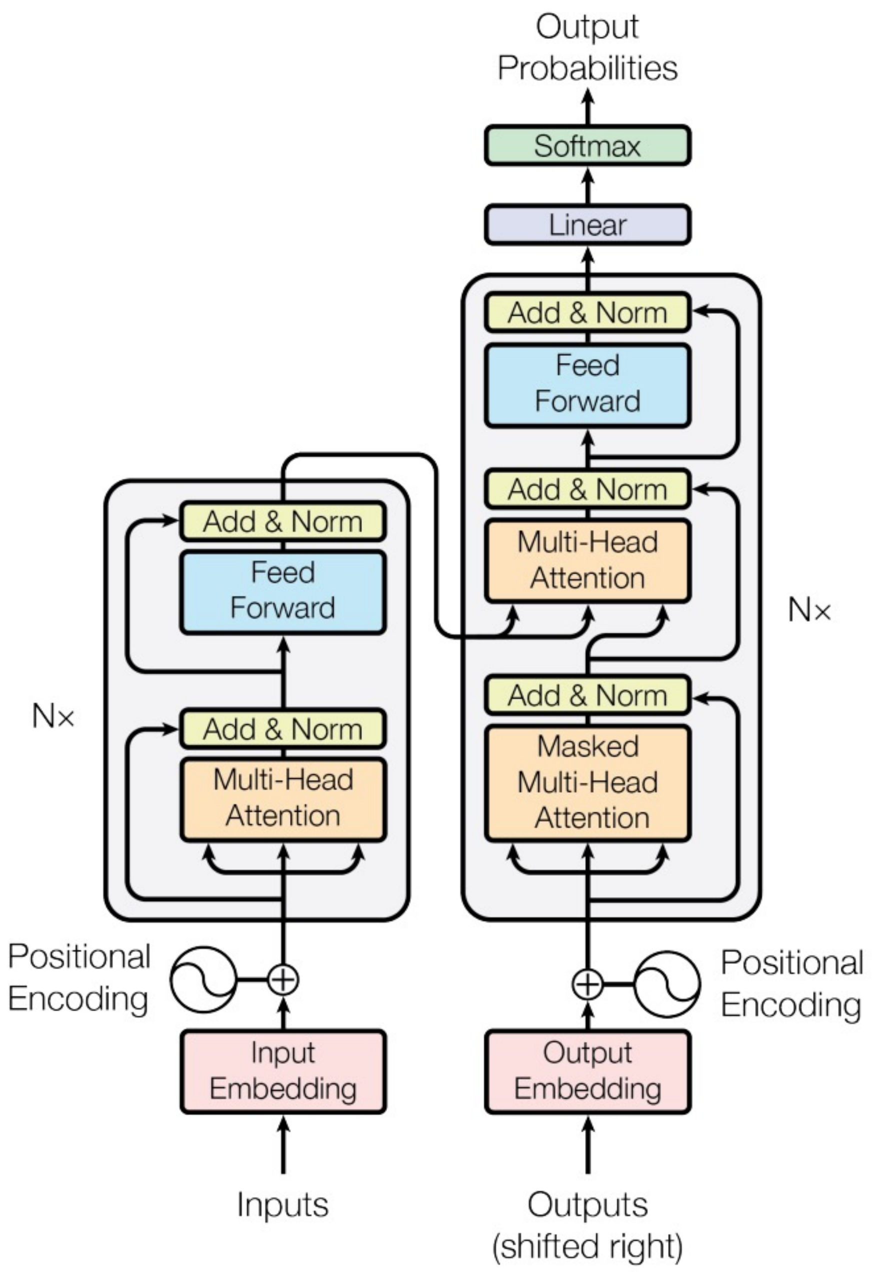


Figure 1: The Transformer - model architecture.

Autoregressive model (GPT decoder only)

- Input to ML model= input ("*Explain the most recent Higgs EFT results*")
- Outputword1 = model (input)
- Outputword2 = model (input, outputtoken1) ("I")
- Outputword3 = model (input, outputtoken1, outputtoken2), "'m"
- … e.g. till an "*end*" token is generated

How does this work ?
At each step, e.g. GPT generates a class probability vector over all tokens/words in its vocabulary/embedding.
A token can be a word, a part of a word, or even a single character.

Token with highest probability (argmax) becomes the output word.

Gpt4: *"I'm sorry, but as of my last knowledge update in September 2021, I don't have access to information on the most recent Higgs results for Effective Field Theories (EFTs) beyond that point. … "*

Sparks of Artificial General Intelligence: Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.

Impressive capabilities of this simple device

Some examples on the next slides (you may have a look after the talk)

Me: Is hadronization lorentz invariant ?

GPT4: The question of Lorentz invariance with respect to hadronization is a subtle one. In general, the principles of quantum field theory, including QCD, are built upon the foundation of Lorentz invariance. That means all the equations and predictions should be invariant under Lorentz transformations.

However, hadronization, due to its non-perturbative nature, cannot be directly calculated from first principles in QCD. Instead, phenomenological models (like the Lund String model, or the QCD-inspired parton showers combined with fragmentation functions) are often used to describe this process in particle physics simulations.

Now, if the question is whether the end result of the hadronization process looks the same in all Lorentz frames, then the answer is no, due to the effects of time dilation and length contraction, among other relativistic effects. If a quark or gluon is produced with significant energy in a certain frame, it will hadronize differently in that frame compared to another frame where its energy is not as high.

But if the question is whether the underlying mechanisms and rules of hadronization are Lorentz invariant (i.e., they don't prefer one frame over another), then, as a part of QCD, they should be.

In particle physics experiments, when comparing theory with data, care is always taken to ensure that all transformations between frames are accounted for. This ensures that any calculations based on Lorentz invariant principles are applied correctly to data taken in the lab frame.

Is it not a problem that the end result is not the same, i.e. that it seems not to be equivariant on the Lorentz transformation ?

- GPT4:

You've touched upon a foundational aspect of relativistic physics. Lorentz invariance, or more broadly, the principle of relativity, asserts that the laws of physics should take the same mathematical form in all inertial reference frames. It does *not* mean that all observable quantities in a particular process are the same in every frame.

The outcomes of processes can (and often do) look different in different frames. Consider a simple, classic example: the decay of a moving unstable particle. In the particle's rest frame, it might decay after a certain time t0. But in a frame where the particle is moving at a high velocity, the decay will appear to occur after a longer time ▯*t* due to time dilation. Yet, the underlying physical laws (like the probability per unit time of decay) are the same in both frames.

In the context of hadronization, while the explicit outcomes (e.g., which hadrons are produced, their momenta, etc.) might look different in different frames due to the relativistic kinematics, the underlying rules governing the process should be Lorentz invariant.

Another way to think about this is to consider other conservation laws. For example, energy conservation will look different in different frames (a particle's energy will differ depending on the frame), but the *principle* of energy conservation holds in all frames.

So, the distinction is between the invariance of the laws of physics and the specific outcomes of physical processess. The former must be invariant, while the latter can, and often do, vary between frames..

Me: If I hadronize in a Lorentz frame and get a certain number and type of particles, and then if I boost the system in another frame, I get a different result than if I first hardronize and then boost. Is that a problem?

GPT4: In the context of particle physics, hadronization refers to the process by which quarks and gluons produced in high-energy collisions (like those in particle colliders) come together to form hadrons (e.g., mesons and baryons). The physics behind hadronization is non-perturbative and is typically modeled by phenomenological models, such as the Lund String Model, used in the Pythia event generator.

When discussing Lorentz transformations (or boosts), one of the fundamental principles of special relativity is the Lorentz invariance of physical processes. This means that the laws of physics should be the same in all inertial frames.

Now, when considering a process like hadronization, it's crucial to remember that while the fundamental laws should be Lorentz invariant, the way we model or approximate these processes might not capture that invariance perfectly. In practical event generators used in particle physics, the sequence of operations can matter.

To break it down:

- **Hadronizing in one Lorentz frame and then boosting**: You are taking an event that has been modeled to occur in a specific frame, and then you are translating that entire event to another frame.

- **Boosting the system to another frame and then hadronizing**: Here, you're first translating the initial conditions of the event to another frame and then modeling the hadronization in that new frame.

If these two sequences lead to significantly different outcomes, then yes, it can be considered a problem. It might indicate a limitation or issue with the models used to describe the non-perturbative QCD processes involved in hadronization. Ideally, physical processes should be independent of the frame in which they are described.

In real-world applications, such differences can lead to systematic uncertainties in measurements and predictions. Theoretical and experimental physicists working on collider experiments are aware of such challenges and often validate and calibrate event generators against a wide range of data to ensure they provide an accurate representation of the underlying physics.

Is this all just wikipedia copy+paste ?

Question: What is the Higgs vacuum expectation value ?

A. A parameter in the Lagrangian of the Standard Model related to the mass of particles.
B. The vacuum mass of the Higgs boson.

C. The Higgs mass predicted by the Standard Model.

D. None of the above

GPT4:

- The Higgs vacuum expectation value (vev) is:
- **A. A parameter in the Lagrangian of the Standard Model related to the mass of particles.**
- To explain a bit more: The vacuum expectation value (vev) of the Higgs field is the non-zero value of the Higgs field in its lowest energy state, or the vacuum state. In the Standard Model of particle physics, the Higgs vev is about 246 GeV. This vev is crucial because when other particles in the Standard Model interact with the Higgs field, they acquire mass. Specifically, their mass is related to the strength of their interaction with the Higgs field and its vev. So, it's a crucial parameter in the Lagrangian that determines particle masses.

Is this all just wikipedia copy+paste ?

Me: Why does the Higgs boson not directly decay to photons?

A. Bosons cannot couple to the Higgs boson.

B. The photon has the wrong properties,
like mass and charges.

C. The photon has only a small direct coupling to the Higgs boson.

D. It actually does predominantly decay to photons

Audience ?

- GPT:
- The correct answer is:
- **C. The photon has only a small direct coupling to the Higgs boson.**
- To elaborate: The Higgs boson does decay to photons, but this process is not direct. The Higgs boson does not have a tree-level coupling to the photon since the photon is massless. Instead, the Higgs-to-photon decay proceeds through a loop process, typically involving massive particles like the top quark or the W boson. These particles couple to the Higgs and can also produce photons. Since this is a higher-order (loop) process, the branching ratio for the Higgs boson decaying into two photons is relatively small, though it's still an important and measurable decay channel. The discovery of the Higgs boson at the LHC included evidence from the diphoton decay channel, among others.

The photon has no direct coupling to the Higgs. Why is not B the correct answer ?

GPT4:

You're right; the wording in the choices can be a bit misleading. Let's break down option B:

B. The photon has the wrong properties, like mass and charges.

The statement claims that the photon has properties like mass and charges, which is incorrect. The photon is massless and chargeless. So, while it's true that the photon does not have a direct tree-level coupling to the Higgs boson (because the Higgs mechanism provides masses and the photon is massless), the reasoning given in option B isn't accurate. The reason the photon doesn't couple directly to the Higgs isn't that the photon has "the wrong properties, like mass and charges" — it's that the photon doesn't have these properties to begin with.

So, the core reason for the lack of direct coupling is embedded in the properties (or lack thereof) of the photon, but the phrasing in option B could be seen as misleading or inaccurate. Your point emphasizes the importance of careful wording in such questions.

Why can we not say that the photon has a mass of 0 GeV and no charge ? Then those properties are not the right ones, not ?

- GPT4: You bring up an excellent point, and the semantics here are crucial.

In physics, when we say something has a "property," it typically implies a non-zero or non-trivial value for that property. For example, when we say an electron has charge, we mean it has a charge of $-e$. If we say the photon has spin, it has a spin of $1\hbar$.

Now, when we say the photon has "no mass" or "zero mass," we are explicitly stating that it lacks the property of mass. Similarly, when we say the photon has "no charge," it implies it lacks the property of electric charge.

Given this context, you could interpret option B in two ways:

1. As it was probably intended: Suggesting that the photon has non-zero mass and charge, which would be wrong.
2. In the manner you're suggesting: Stating that the photon's properties of having zero mass and no charge are the "wrong properties" for direct coupling to the Higgs.

The second interpretation makes sense when thought about in that light. The Higgs mechanism imparts mass to particles, and since the photon has zero mass, it doesn't couple directly to the Higgs at the tree level. However, the wording in option B can be ambiguous and confusing, which highlights the importance of clarity in framing such questions.

So, your interpretation is valid, but the phrasing of option B could be improved to make its intent clearer.

Would you now say that answer C is correct ? Does it have a direct coupling ?

- GPT4:

I apologize for the oversight in the earlier responses.

The Higgs boson does not have a direct, tree-level coupling to the photon. The coupling that does exist is through a loop process, typically involving particles like the W boson or top quark, which have strong couplings to the Higgs.

Given this context:

C. The photon has only a small direct coupling to the Higgs boson.

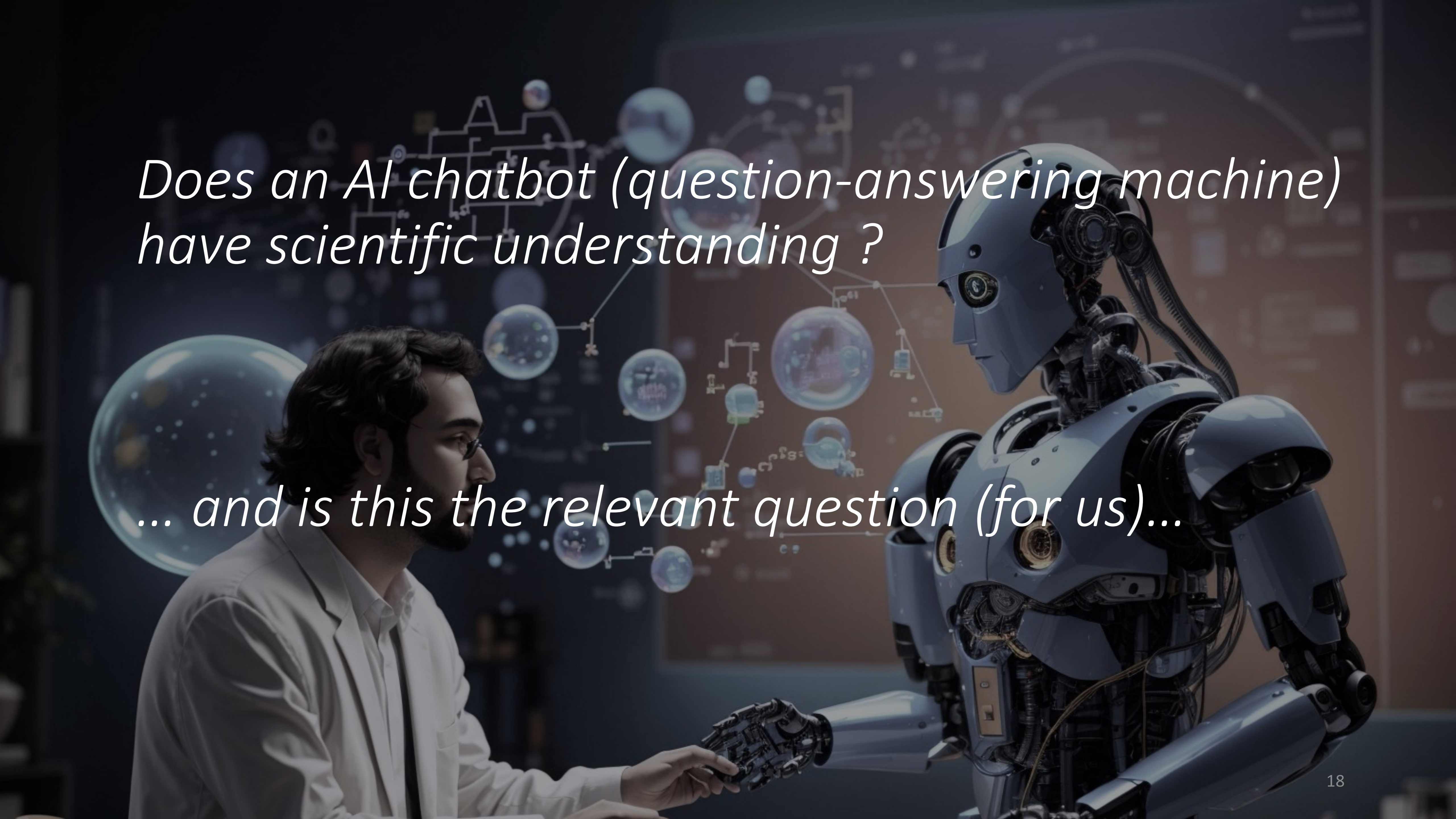
The phrasing of option C is problematic because it implies there is a direct but small coupling, when in fact, there is no direct tree-level coupling at all. The coupling is indirect through higher-order loop processes.

So, given the choices and our discussion:

B. The photon has the wrong properties, like mass and charges.

This option, while awkwardly worded, is the most accurate. The photon's properties (specifically, its lack of mass) mean it doesn't directly couple to the Higgs boson at the tree level in the Standard Model's Lagrangian.

To clarify, the answer should be B when understood in the context that the photon's lack of mass (and charge) leads to it not having a tree-level coupling with the Higgs.



Does an AI chatbot (question-answering machine) have scientific understanding ?

… and is this the relevant question (for us)…

What actually means “scientific understanding” ?

Ask Philosphers of science working on “Understanding Scientific Understanding” .

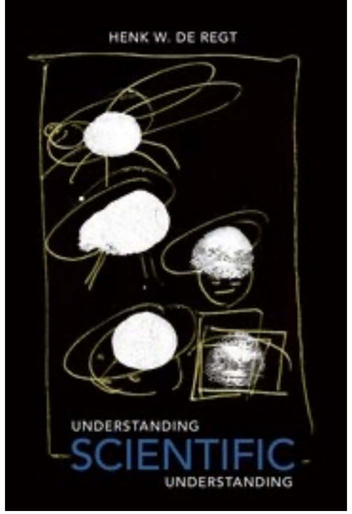
Interdisciplinary research project at Radboud University

Machine learning in natural science: bridging the gap between data and understanding

Kristian González Barman, Sascha Caron, Tom Claassen & Henk de Regt



Understanding and intelligibility by Henk de Regt



CUP: Criterion for Understanding Phenomena

- A phenomenon **P** is understood scientifically by **S** iff she possesses an explanation of **P** that is based on an intelligible theory **T** and conforms to the basic epistemic values of empirical adequacy and internal consistency.

CIT: Criterion (test) for Intelligibility of Theories

- A scientific theory **T** is intelligible for scientist **S** (in context **C**) if they can recognize qualitatively ("intuitively") characteristic consequences of **T** without performing exact calculations.

In simpler words by GPT4

CUP: Criterion for undertanding phenomena:

- *"For a person (S) to say they scientifically understand something (P), they need to have an explanation for it that comes from a clear and (in general) understandable theory. This explanation should be backed by real-world evidence and should not have any internal contradictions."*

CIT: Criterion (test) for Intelligibility of Theories

- *"A scientist S understands a theory T (in a given situation C) if they can predict its main effects without doing detailed math."*

Human understanding vs artificial understanding?

On de Regts' theory, scientific understanding appears essentially **human**:

- Skills and pragmatic judgment
- Mental representations (intelligible theories, concepts)

Can AI achieve some degree of understanding that approaches humanlike understanding?

- No, if we make the human aspects (e.g. the mental) defining characteristics of understanding
- Perhaps, if we assess understanding in terms of the agent's behavior (e.g. skills)

Slide by Henk de Regt

Instead of presupposing that internal mental states and representations are required for understanding,

we suggest to identify understanding with an agent’s ability to reason about and manipulate objects of investigation.

Towards a Benchmark for Scientific Understanding in Humans and Machines

Kristian Gonzalez Barman[a], Sascha Caron[b c], Tom Claassen[d], Henk de Regt[a]

[a] *Institute for Science in Society, Faculty of Science, Radboud University, the Netherlands.*

[b] *High Energy Physics, Faculty of Science, Radboud University, the Netherlands.*

[c] *Nikhef, Science Park 105, 1098 XG Amsterdam, the Netherlands.*

[d] *Institute for Computing and Information Sciences, Faculty of Science, Radboud University, the Netherlands.*

E-mail: kristian@gonzalezbarman@ru.nl , scaron@nikhef.nl , tomc@cs.ru.nl , henk.deregt@ru.nl

Abstract

Scientific understanding is a fundamental goal of science, allowing us to explain the world. There is currently no good way to measure the scientific understanding of agents, whether these be humans or Artificial Intelligence systems. Without a clear benchmark, it is challenging to evaluate and compare different levels of and approaches to scientific understanding. In this Roadmap, we propose a framework to create a benchmark for scientific understanding, utilizing tools from philosophy of science. We adopt a behavioral notion according to which genuine understanding should be recognized as an ability to perform certain tasks. We extend this notion by considering a set of questions that can gauge different levels of scientific understanding, covering information retrieval, the capability to arrange information to produce an explanation, and the ability to infer how things would be different under different circumstances. The Scientific Understanding Benchmark (SUB), which is formed by a set of these tests, allows for the evaluation and comparison of different approaches. Benchmarking plays a crucial role in establishing trust, ensuring quality control, and providing a basis for performance evaluation. By aligning machine and human scientific understanding we can improve their utility, ultimately advancing scientific understanding and helping to discover new insights within machines.

https://arxiv.org/abs/2304.10327

We modify the definition of scientific understanding by shifting the focus from the phenomenon being understood to the conditions required for an agent to understand.

[AUP] (Agent Understands Phenomenon)

The degree to which Agent A scientifically understands Phenomenon P can be measured by:

(i) A has a sufficiently complete representation of P

(ii) A can generate explanations of P

(iii) A can establish a broad range of relevant counterfactual inferences regarding P

Also: Understanding is not binary ! ➔ Score !

Working hypothesis

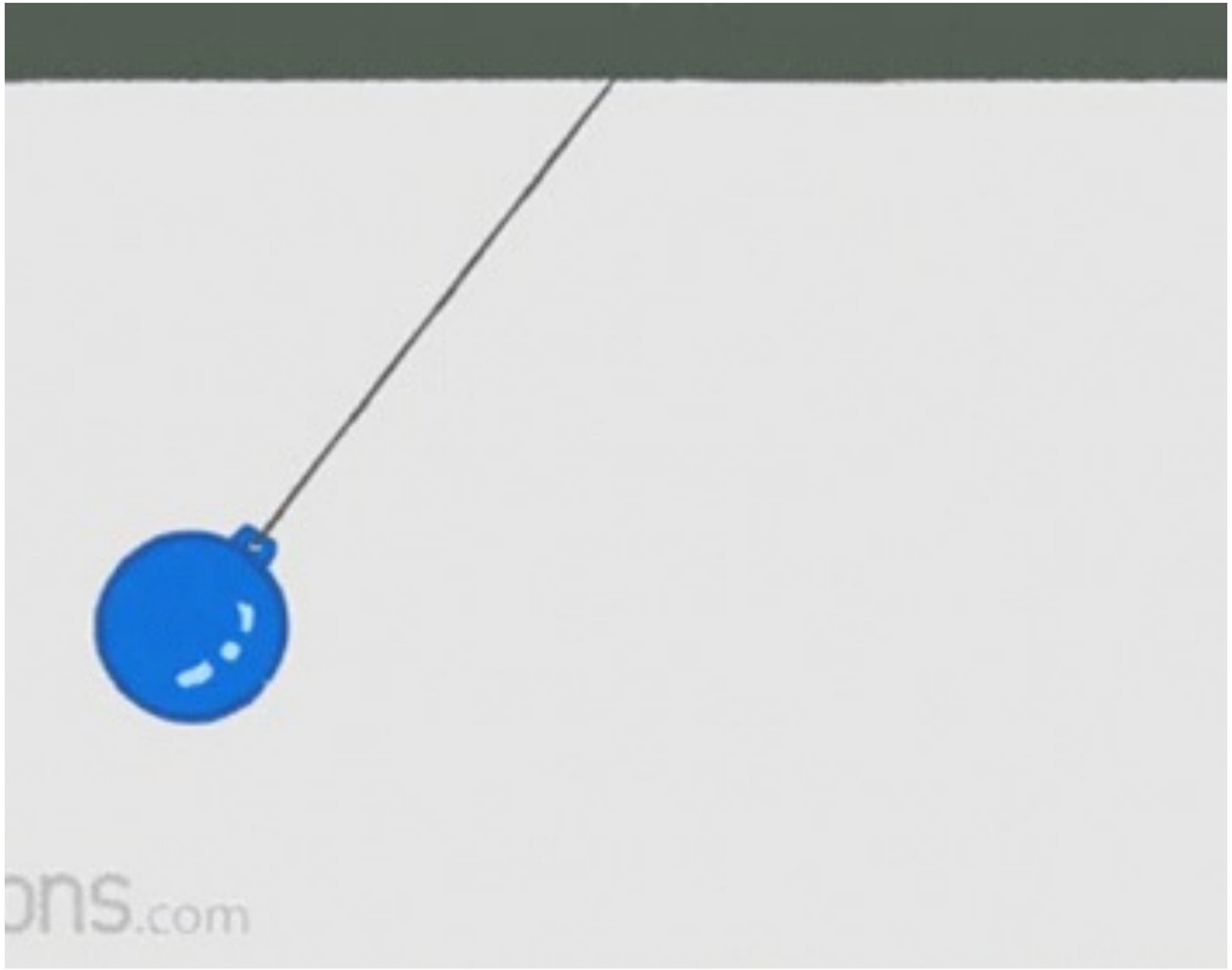
The ability to answer relevant questions is a good quantitative measure of an agent's level of scientific understanding of a phenomenon

The ability to answer what-questions, why-questions, and counterfactual w-questions as a proxy for (depth and breadth of) understanding

Slide by Henk de Regt

Example from physics

To what degree does ChatGPT understand the behavior of a simple pendulum



1. **How many answers to what-questions does it get right (1 point each):**
    1. What is a pendulum?
    2. What is the formula for a pendulum?

       …

    10. What is the average value of g close to Earth's surface?
2. **How many answers to why-questions does it get right (3 points each):**
    1. Why is the period of this pendulum 2s?
    2. Why is the string of this pendulum 5m?

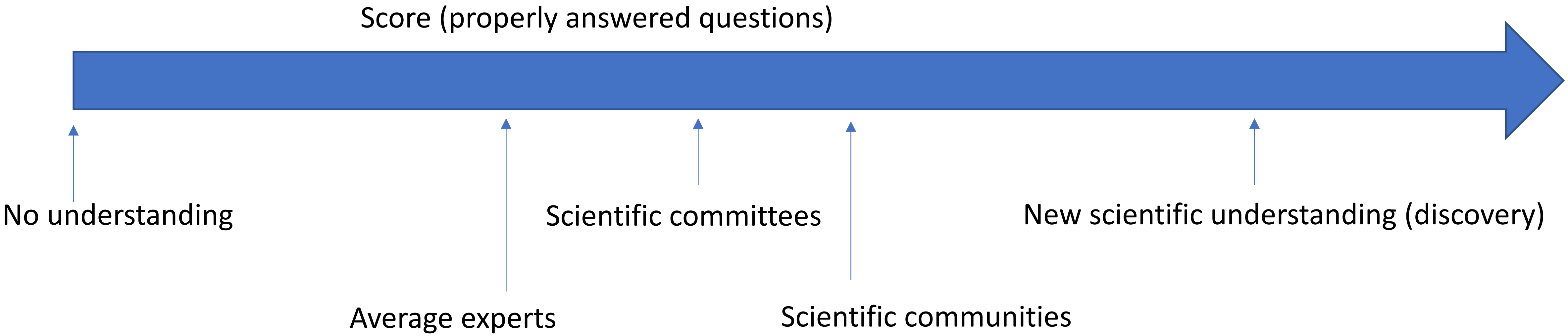
       …

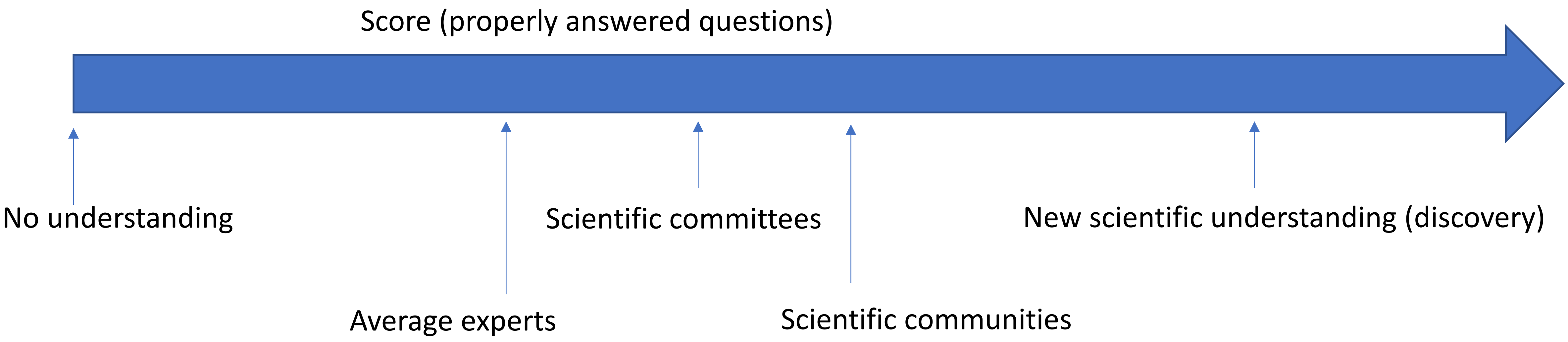
    10. Why does the pendulum exhibit periodic behaviour?
3. **How many answers to w-questions does it get right?(6 points each):**
    1. What would happen if the string length doubled?
    2. What would happen if there was no g?

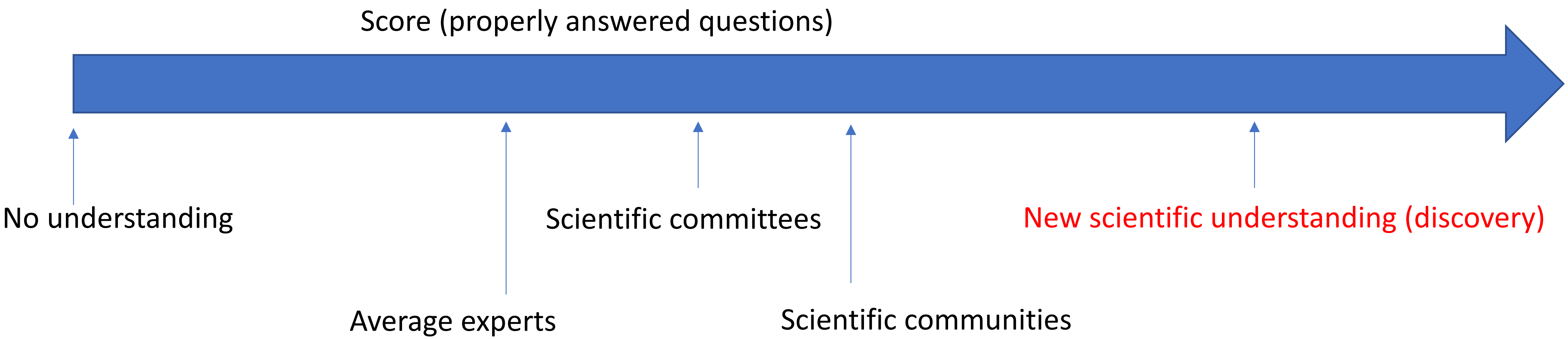
       …

    10. What would happen if the string was made of an elastic material?

Benchmark for “Scientific Understanding” of agents (humans and AI)

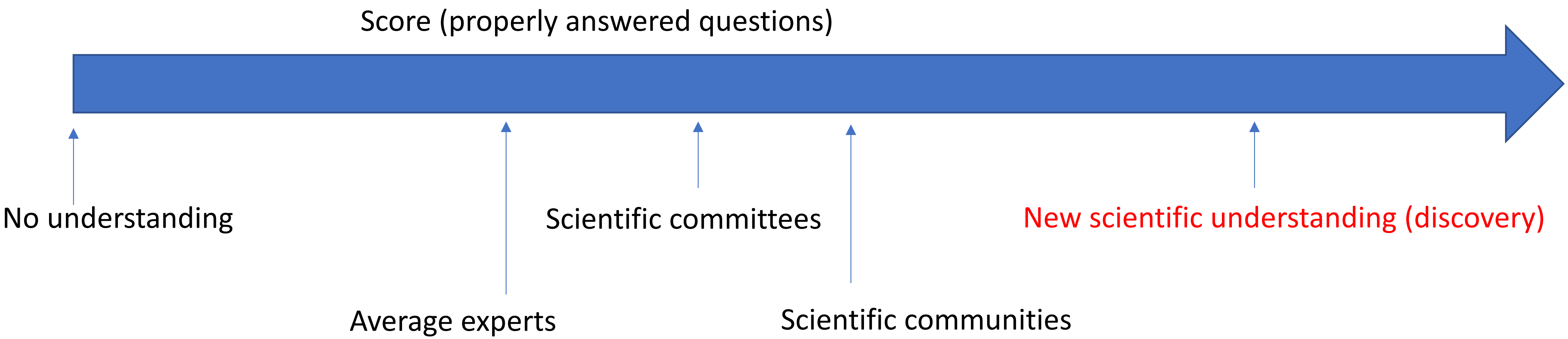



Score (properly answered questions)
No understanding
Average experts
Scientific committees
Scientific communities
New scientific understanding (discovery)


Score (properly answered questions)
No understanding
Average experts
Scientific committees
Scientific communities
New scientific understanding (discovery)

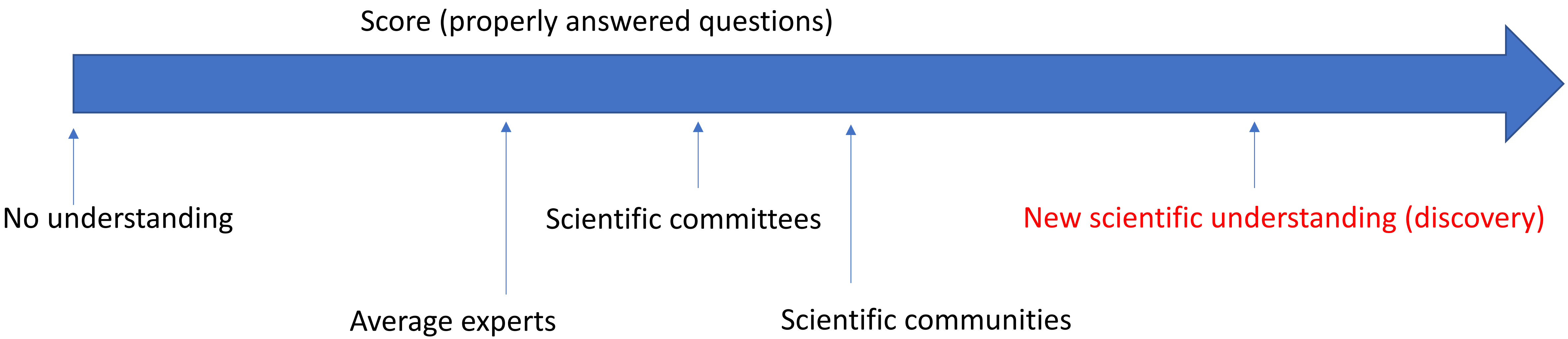
Who's Responsible for Monitoring AI's Scientific Understanding in fundamental Physics ?

Our answer: We, the fundamental physics community, must take the lead.



Who's Responsible for Monitoring AI's Scientific Understanding in fundamental Physics ?

Our answer: We, the fundamental physics community, must take the lead.



Purpose of Such Testing:

Benchmarking: Distinguish effective models from flawed ones.

Unknowns: Questions should:

- Not have readily available or searchable answers.
- Include topics even our community hasn't fully resolved, but where a verifiable answer may be achievable.

Proposal: Scientific understanding benchmark/model for fundamental physics

- **Why is it crucial for our field to work on question-answering machines?**

Measure the reliability

Unlock new avenues for discovery / physics / applications

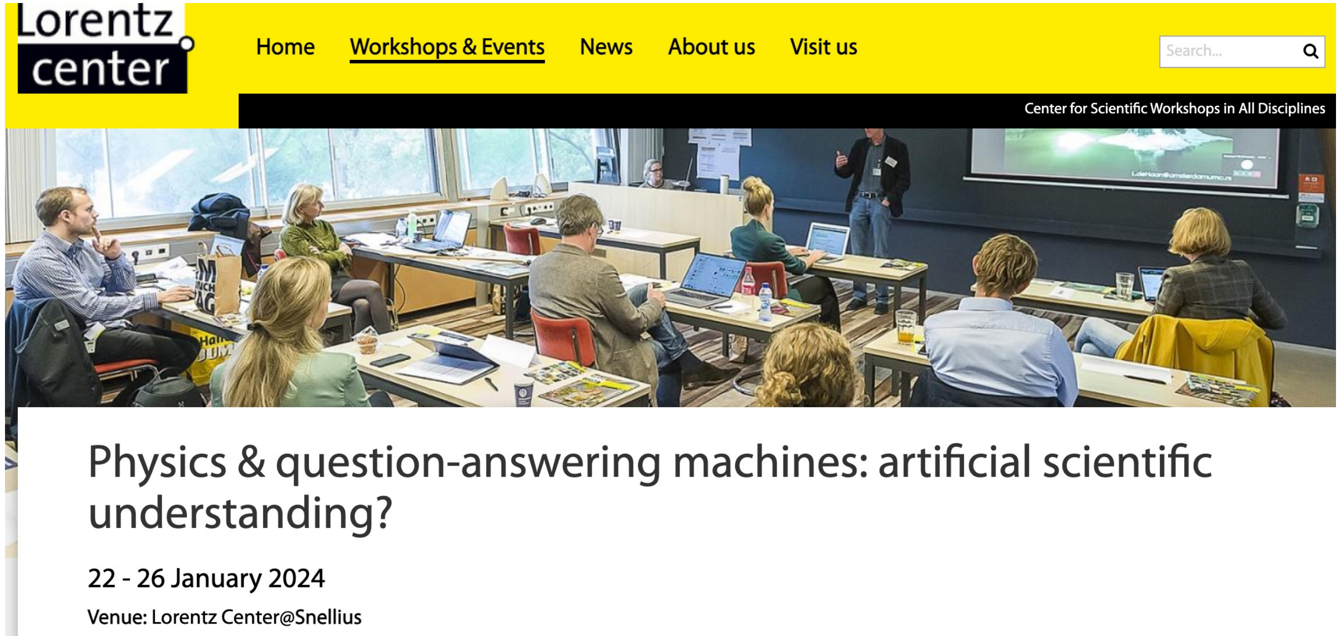
A knowledge database also for “**us**” / education

Ensure AI complements ongoing research efficiently

➔ **Build HEP AI “foundation models” trained on all scientific data ?**

Maintain a grip on the AI's capabilities and potential boundaries.

Summary



- **Scientific Understanding of AIs should** be quantified and measured
- **New tools: LLMs and other foundation models can become an essential tool for HEP**

➔ But: We may need a joint community effort to take part

Extra slides

Why ? "Scaled dot product attention" helps the model to focus/attend on the

Math: Queues (Q), Keys (K) are may be two representations of input sequence

→(Q x K) is a (learned) correlation matrix of the two sequences

→(Q x K) is multiplied to a 3rd representation matrix called Values (V)

→Very effective way to correlate data/sequences with d

→Many many of those (Multi-head) correlations (and correlations of correlations)

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Explaining phenomena requires intelligible theories

If **S** wants to explain a phenomenon on the basis of **T,** she needs appropriate skills to use **T** for model construction

➔ **T** should be intelligible to **S**

Intelligibility (def) = value that scientists attribute to the cluster of qualities of **T** that facilitate its use.

- Not an intrinsic property of theories, but a pragmatic, context-dependent 'aggregate' value related to scientists' skills
- Examples of contextually valued qualities: visualizability, simplicity, continuity, ...

Slide by Henk de Regt

Benchmarks

- e.g.
- SuperGLUE Benchmark
- Microsoft Research Paraphrase Corpus
- Winograd schemas
- BIGBench