FOUNDATION MODELS FOR PHYSICS

Nesar Ramachandra
nramachandra@anl.gov
CPS Division, Argonne National Laboratory

U.S. DEPARTMENT OF ENERGY | Office of Science

Argonne NATIONAL LABORATORY

INTRODUCTION

- What are Foundation Models (FMs)? What makes them particularly interesting or powerful?
- What does it take to train a Large Language Model (LLM)?
- What are restrictive set of problems can we deal with*? How do we repurpose LLMs or FMs into our scientific problems?

*Disclaimer: Some of the opinions are my own.

HOW ARE FOUNDATION MODELS DIFFERENT?

Argonne NATIONAL LABORATORY

Traditional AI:

- *Parameters:* O(1e3-1e8)
- *Data:* Specialized, annotated
- *Flexibility:* Meant for specific tasks.
- *Training Approach:* Supervised, unsupervised, or transfer learning.
- *Applications:* Suited for narrowly defined problems with clear objectives.

Foundation models:

- *Parameters:* O(1e8 to 1e11)
- *Data:* Extensive, datasets from various domains
- *Flexibility:* Highly versatile, can be fine-tuned different tasks.
- *Training Approach:* Generic pre-training followed by alignments
- *Applications:* Suitable for multi-faceted problems requiring nuanced understanding

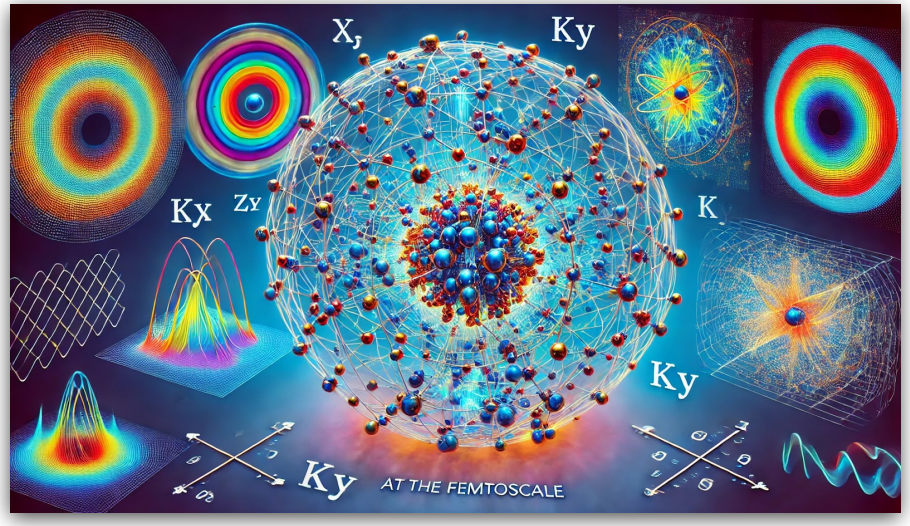


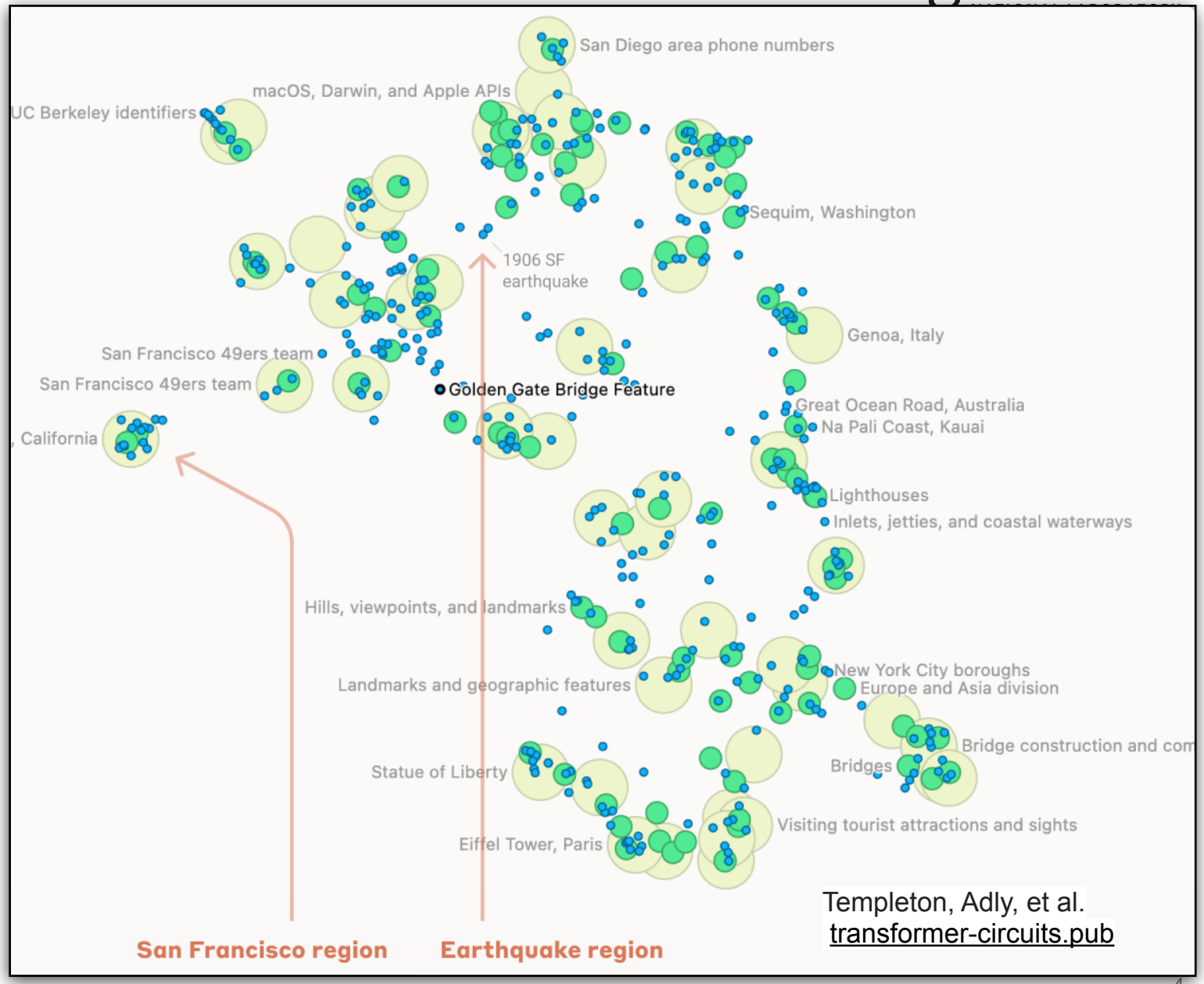
Image credits: Emil Constantinescu & ChatGPT

Gemini

code llama

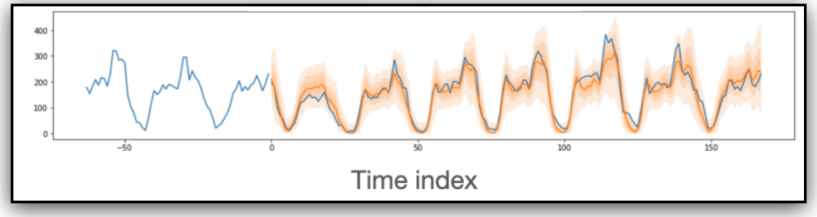
DO FOUNDATION MODELS GENERALIZE?

- Majority of the models are language based.
- Emergent behaviors seen with larger models.
- Investments from big-tech, big GPU clusters being utilized.
- Questions about whether this is the path towards AGI, whether there is a reasoning/ thinking involved at inference.
- Regardless, FMs are technological paradigm shifts, and a highly useful knowledge base.



Templeton, Adly, et al.
transformer-circuits.pub

BUILDING BLOCKS OF FOUNDATION MODELS



- Transformer blocks are the fundamental blocks, self-attention is essential functionality:
  - From "**Attention is all you need:** https://arxiv.org/abs/1706.03762"
- Successor to Recurrent Neural Networks (RNNs), designed to predict next occurrence in a time series.
- Self attention is key to process large amount of data at once.
- Associated 'technical' details: Tokens, Token embeddings, Position encoding, Vocabulary, Context, Context window lengths, Multi-head attention.

Sequence of tokens: $U = \{u_1, u_2, u_3, \ldots, u_N\}$

Minimize: $L(U) = \sum_{i=1}^{N} log(p(u_i \mid u_{i-k}, \ldots, u_{i-1}, \Theta))$

(Conditional probability of i-th token given k preceding tokens)

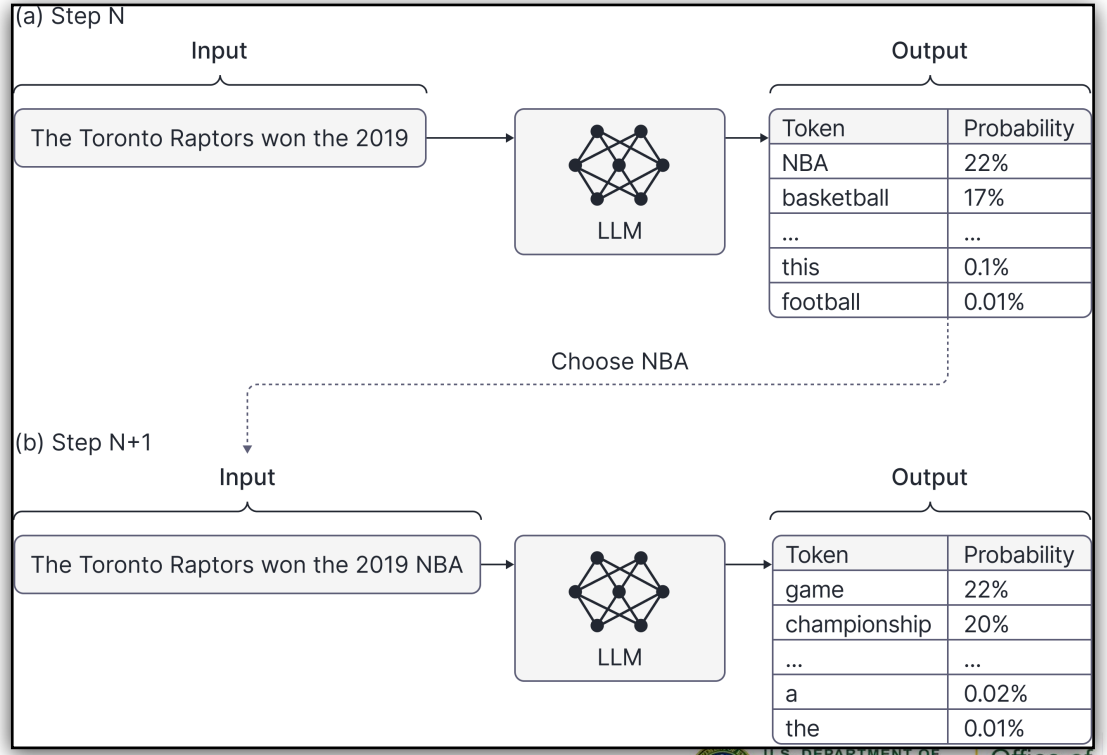


Image source:borealisai

STAGES IN TRAINING AN LLM

- **Data preparation**
  - Both labeled and unlabeled datasets.
- **Pre-training**
  - Computationally expensive
- **Alignment**
  - Smaller task, includes fine-tuning or instruction tuning.
- **Reinforcement learning from Human Feedback (RLHF)**
  - Safety guard-rails
- **Benchmarking and validation**
- **Deployment and API**

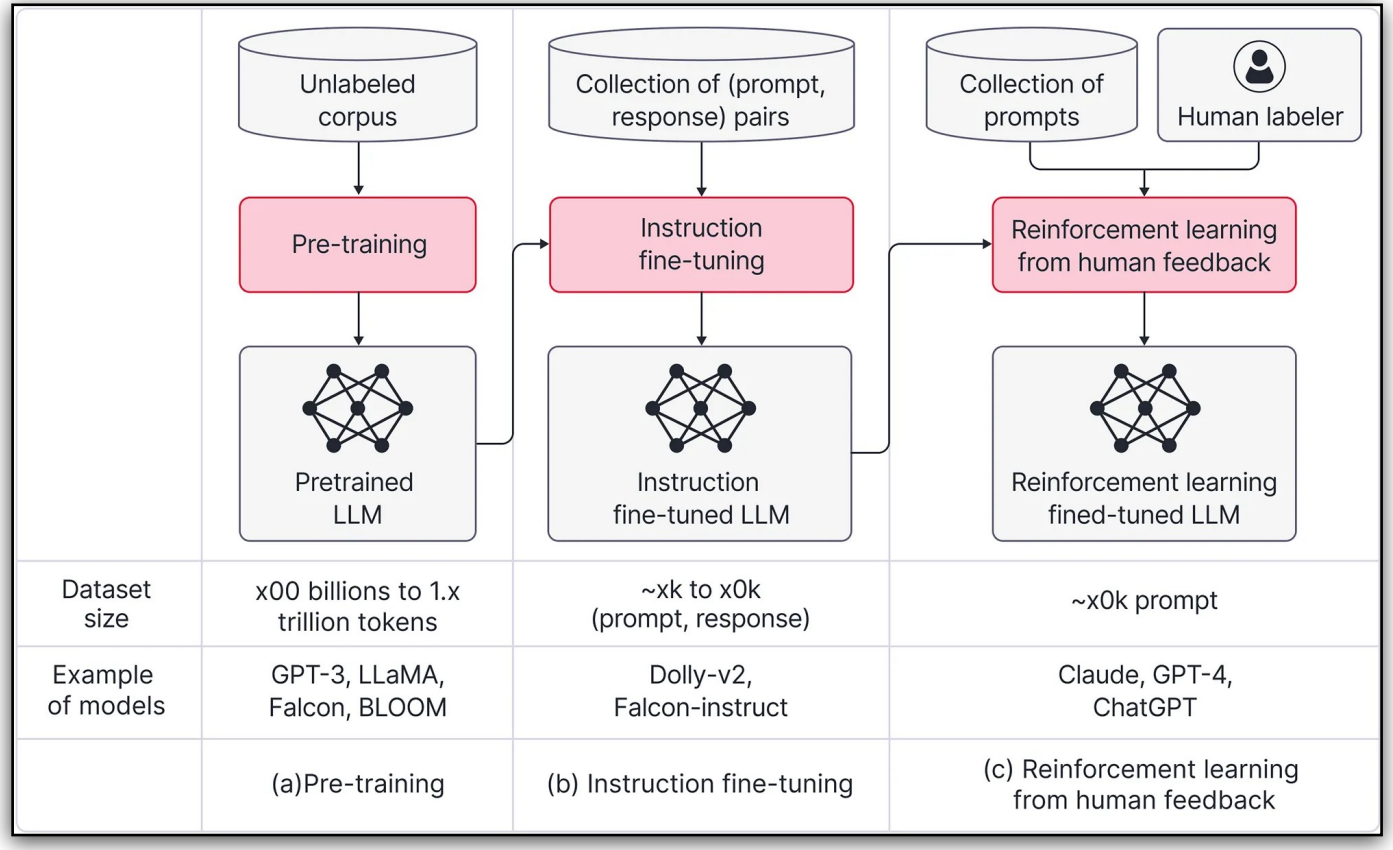


Image source:borealisai

LLM-ASSISTED RESEARCH: TEXT ONLY

- **Direct usage**
  - Text-based queries for knowledge search.
  - Prompt engineering
- **Fine-tuning**
  - Smaller, specific scientific datasets (Ex: AstroLLaMa)
- **Full model building**
  - Expensive pre-training from scratch (Ex: AuroraGPT)
    - Scientific datasets
    - Scientifically rigorous benchmarks

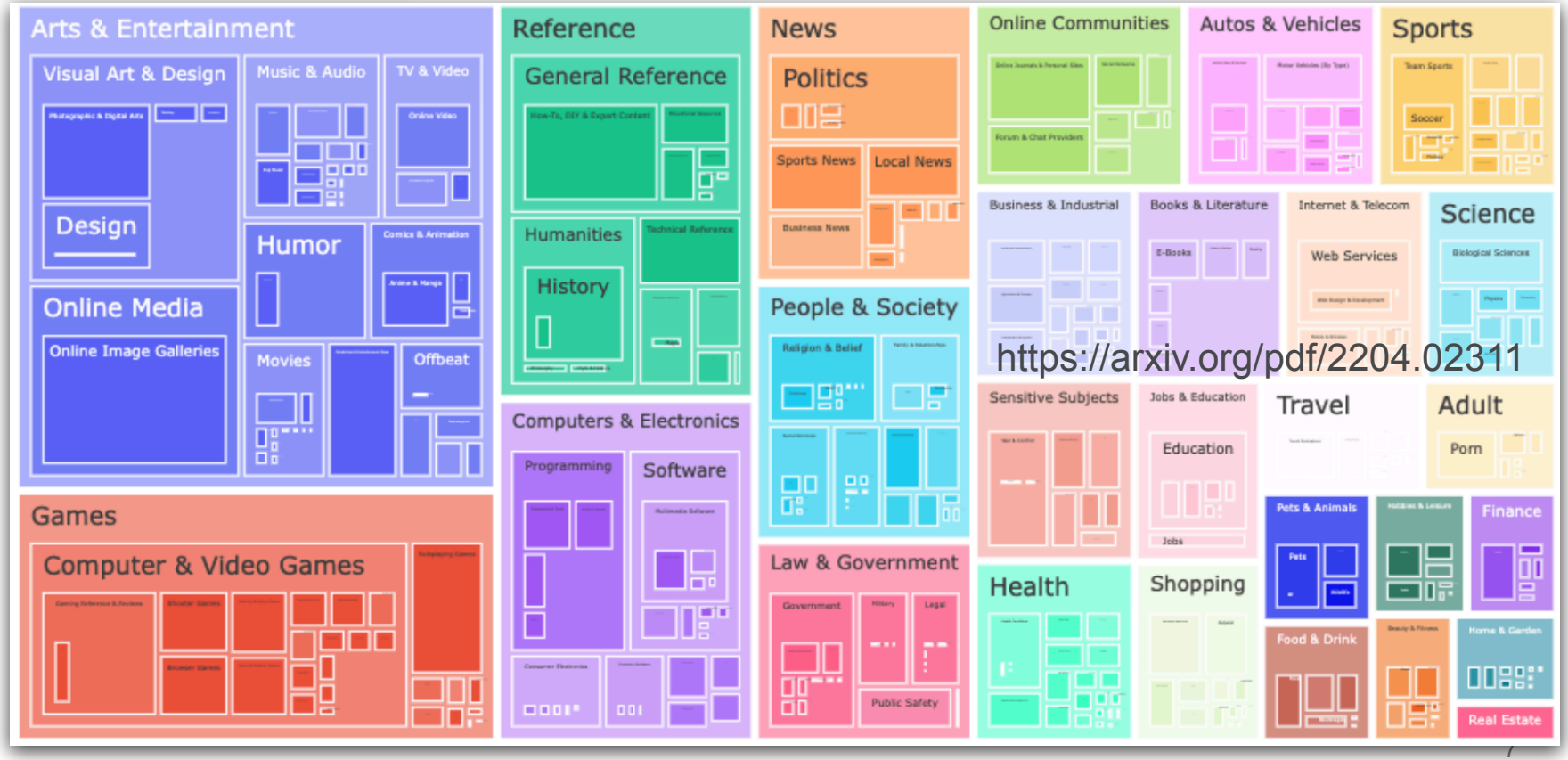
Completed by LLaMA-2

The Magellanic Stream (MS) - an enormous ribbon of gas spanning 140◦ of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We find a stellar stream in the MS, with a 6D velocity of 100 km s−1, and a distance of 100 kpc. The stream is 100 pc wide and 1000 pc long, and is consistent with the MS. The stream is 1000 times more massive than the Magellanic Clouds, and is likely to be a tidal stream from the Large Magellanic Cloud.

Completed by AstroLLaMA

The Magellanic Stream (MS) - an enormous ribbon of gas spanning 140◦ of the southern sky trailing the Magellanic Clouds - has been exquisitely mapped in the five decades since its discovery. However, despite concerted efforts, no stellar counterpart to the MS has been conclusively identified. This stellar stream would reveal the distance and 6D kinematics of the MS, constraining its formation and the past orbital history of the Clouds. We 1) present a new method to select MS candidate stars from wide field optical imaging, 2) apply this method to the MS, 3) compare the results to existing catalogs, and 4) identify the most promising MS candidate stars. We apply our method to 2547 stars in the Gaia EDR3 catalog that have Gaia proper motions consistent with the MS and use these as a pilot sample to optimize our method. We then apply our method to the entire Gaia EDR3 catalog to identify 46,000 stars that are candidates for the MS, and we use the Gaia-ESO Survey to identify 13,000 of these stars that are likely to be MS candidates. We identify 266 stars with (G-K) colours and [Fe/H] that are consistent with previous estimates of the MS. These stars have proper motions and radial velocities that are consistent with the MS, and we place them on a 3D map of the MS. We find a previously unknown stellar stream, which we call the Southeastern Stream (SES), in the direction of the Clouds and MS. The SES has an M31 distance of 32.2+/-0.5 kpc and a 6D velocity of 136+/-10 km/s. The SES is about 20 times brighter in stars than the MS and about 40 times brighter in stars than the SMC. The SES is also 2.7 times more metal-rich than the MS, which suggests that it is a separate population.

https://arxiv.org/pdf/2309.06126



https://arxiv.org/pdf/2204.02311

U.S. DEPARTMENT OF ENERGY | Office of Science

PHYSICS BENCHMARKS FOR LANGUAGE MODELS

- Multiple choice questions and Free-form questions can be used to evaluate and compare models
- Compared to biology/chemistry, available LLM benchmarks for physics/astronomy is inadequate
  - Both in terms of quantity and quality

Where do most short-period comets come from and how do we know?
Say the pupil of your eye has a diameter of 5 mm and you have a…
Why isn't there a planet where the asteroid belt is located?
Why is Mars red?

mmlu-astronomy

Electromagnetic radiation provides a means to probe aspects of the physical universe. Which of the following statements regarding radiation spectra is NOT correct?	["Lines in the infrared, visible, and ultraviolet regions of the spectrum reveal primarily the nuclear structure of the sample.", "The wavelengths identified in an absorption spectrum of an element are among those in its emission spectrum.", "Absorption spectra can be used to determine which elements are present in distant stars.", "Spectral analysis can be used to identify the composition of galactic dust."]
One end of a horizontal, massless spring is attached to a wall. A mass of 0.30 kg is attached to the other end of the spring and rests on a table. The mass is displaced…	["0.24 mJ", "0.38 mJ", "0.48 mJ", "0.75 mJ"]

mmlu-college_physics-original-neg

Things that are equivalent according to the equivalence principle are	["space and time.", "a traveling twin and a stay-at-home twin.", "gravity and acceleration.", "mass and energy."]
Colors in a soap bubble result from light	["converted to a different frequency", "deflection", "interference", "polarization"]

mmlu-conceptual_physics-dev

BENCHMARKING FOR AURORA-GPT

- Benchmark development may be crucial for a Science-focussed GPT comparison with existing LLMs.
- Benchmarking team at the Aurora-GPT collaboration has released a web-form to collect science questions of interest — with real-time evaluation from multiple LLMs.
- Goal is to collect O(1000) questions across scientific fields — **HELP needed! (and potential collaboration opportunities)**
  - High-quality.
  - Should represent what the science community wants out of an LLM.
  - Should not be exposed to current LLMs.

https://web.cels.anl.gov/projects/auroragptquestions/ui/index.html

Trial Q:For the force between quarks, which of the following statements is true?

a) The force follows an inverse square law
b) Approximately an inverse square law but asymptomatically weaker at short distances
c) Approximately an inverse square law but asymptomatically stronger at short distances
d) Approximately an Inverse linear relationship, but goes asymptomatically free at large distances.
e) Asymptomatically free at large distances and small distances, inverse relationship in between.

SEMI-AUTOMATED BENCHMARKS

Argonne NATIONAL LABORATORY

- Existing LLMs can be used to construct questions from a small number of papers (review papers, white papers).
  - Deeper than current MCQs, yet not hyper-specific to papers.
  - Scaling to O(10000) questions is relatively more straightforward.

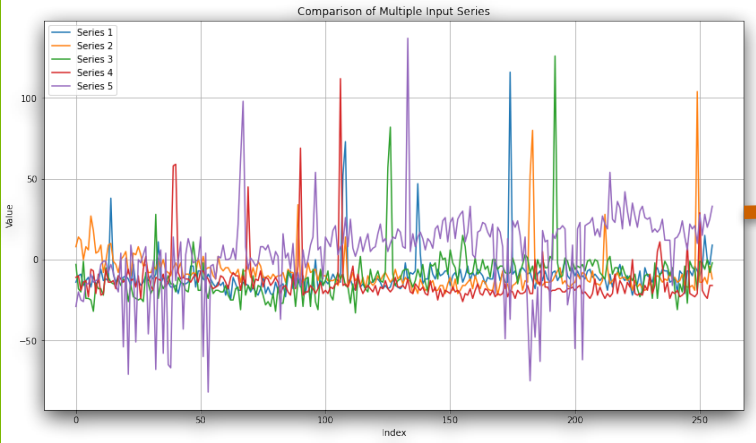

```
"question": "What is the predominant model
    for cosmic ray (CR) transport in the
    Galaxy at energies below about 10^17 eV?",
"distractors": [
    "Convective wind model with uniform
        source distribution",
    "Static halo model with isotropic source
        emission",
    "Leaky-box model with constant escape
        time"
],
"correct_answer": "Diffusion model with
    possible inclusion of convection",
```

```
"question": "If a new particle was discovered
    that interacts with the weak force but
    not the electromagnetic or strong force,
    what could be inferred about its
    properties?",
"distractors": ["It must have a fractional
    electric charge", "It must be massless",
    "It could be a candidate for dark
    matter", "It must be a type of quark"],
"correct_answer": "It could be a candidate
    for dark matter",
```

```
"question": "If gauge bosons were discovered
    to have non-zero rest mass without the
    need for a Higgs mechanism, what would be
    the most likely implication?",
"distractors": ["The existence of a fifth
    fundamental force", "A revision of the
    Standard Model to include mass terms for
    gauge bosons", "The invalidation of gauge
    symmetry", "The discovery of new gauge
    bosons beyond the Standard Model"],
"correct_answer": "A revision of the Standard
    Model to include mass terms for gauge
    bosons",
```

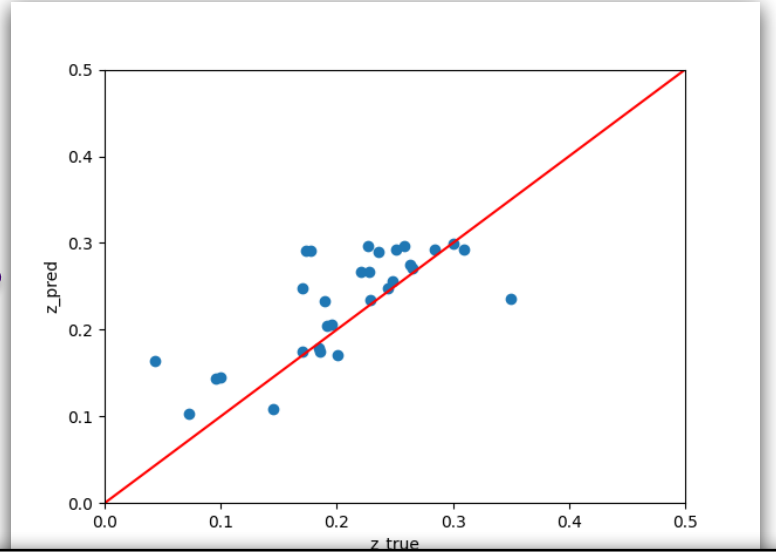
FOUNDATION MODELS BEYOND TEXT: FINE-TUNING

- Missing features in current LLMs:
  - Support for non-text data like scalar numbers, vectors, images, trees, unstructured datasets.
- Can pre-trained LLMs be re-purposed for numerical tasks?
  - It depends!
  - Degradation/forgetting is expected



NR, Yuan-Sen Ting (in prep.)

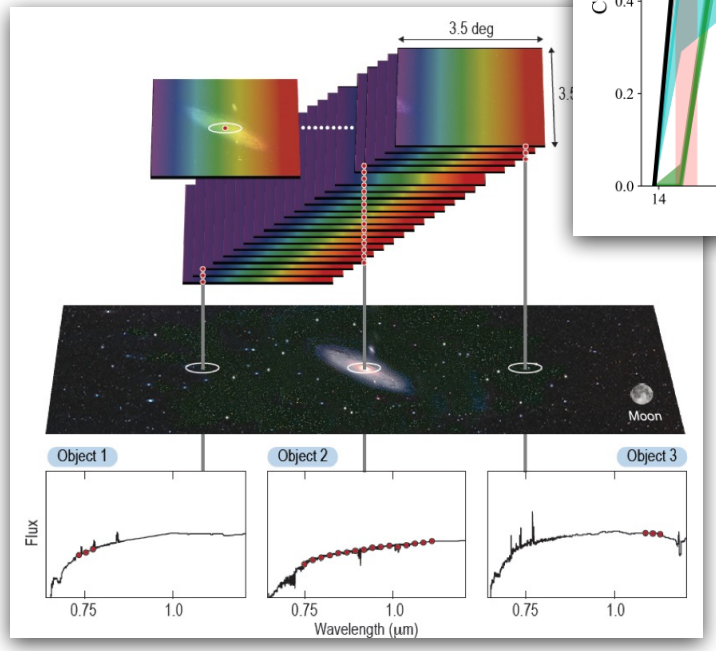
Fine-tuned LLaMa

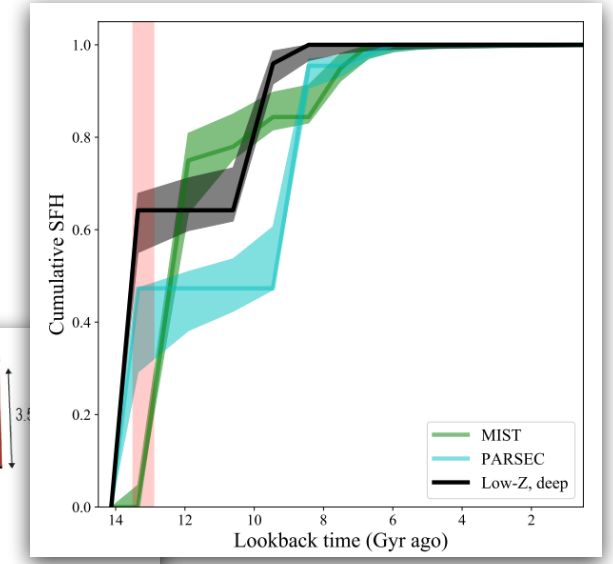



```
Q: What type of galaxy is expected at redshift of z=0.211? List the telescope
    surveys where such a galaxy can be observed.

A galaxy at redshift z=0.211 is expected to be a luminous infrared galaxy
    (LIRG) or an ultraluminous infrared galaxy (ULIRG) due to the fact that
    the universe was in its peak star formation period during that epoch. The
    galaxy is likely to be a merger or a strongly interacting system, and it
    could be a dusty galaxy with a high rate of star formation.

The galaxy can be observed in the following telescope surveys:

1. Sloan Digital Sky Survey (SDSS): The galaxy can be observed in the SDSS-IV
    survey, which covers the sky in five bands (u, g, r, i, and z) and has a
    limiting magnitude of 22.5 in the r-band.
2. Dark Energy Survey (DES): The galaxy can be observed in the DES survey,
    which covers the sky in five bands (g, r, i, z, and Y) and has a limiting
    magnitude of 24.5 in the i-band.
3. Hyper Suprime-Cam (HSC) Survey: The galaxy can be observed in the HSC
    survey, which covers the sky in five bands (g, r, i, z, and Y) and has a
    limiting magnitude of 26.5 in the i-band.
4. Subaru Hyper Suprime-Cam (HSC) Survey: The galaxy can be observed in the
    Subaru HSC survey, which covers the sky in five bands (g, r, i, z, and Y)
    and has a limiting magnitude of 26.5 in the i-band.
```

BEYOND-TEXT: FULL MODEL BUILDING

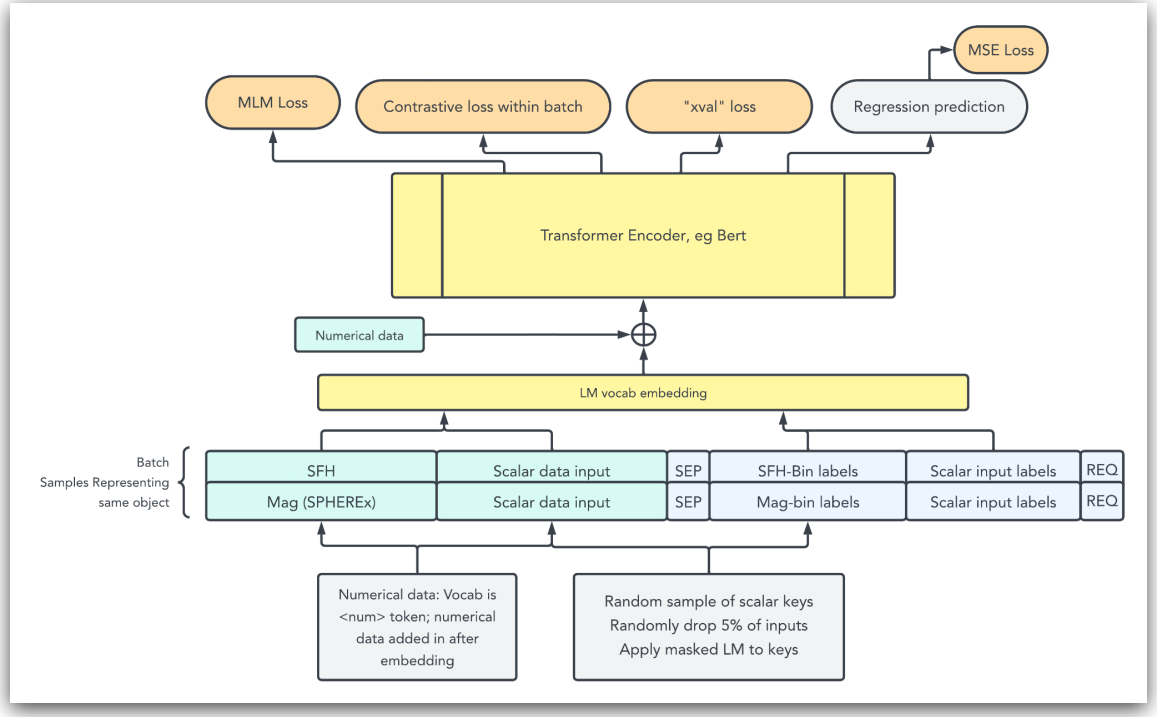
- Foundation models for non-text scientific data formats:
  - For science datasets: Star-foundation models, AstroCLIP, OmniLearn
- Numbers be processed as numbers and not as tokenized representations: xVal
- Multi-modality in object expressions, not data-types.
  - Distinct from industry trend: videos, subtitles, plots and captions, songs and lyrics
- Variety of transformers
  - Vision Transformers, graph transformers
- Latent space modeling using Contrastive learning



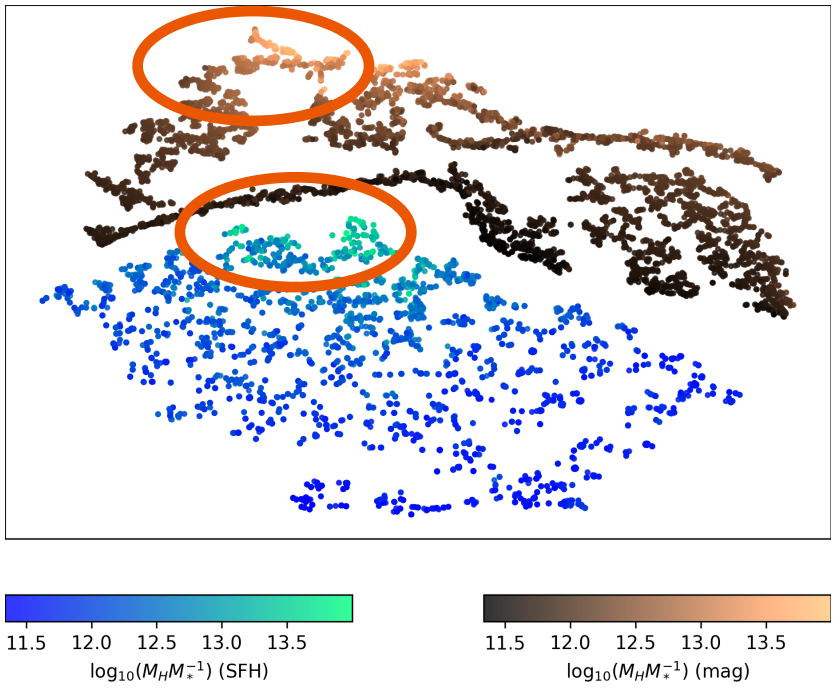


Azton Wells, NR,
Salman Habib, in prep

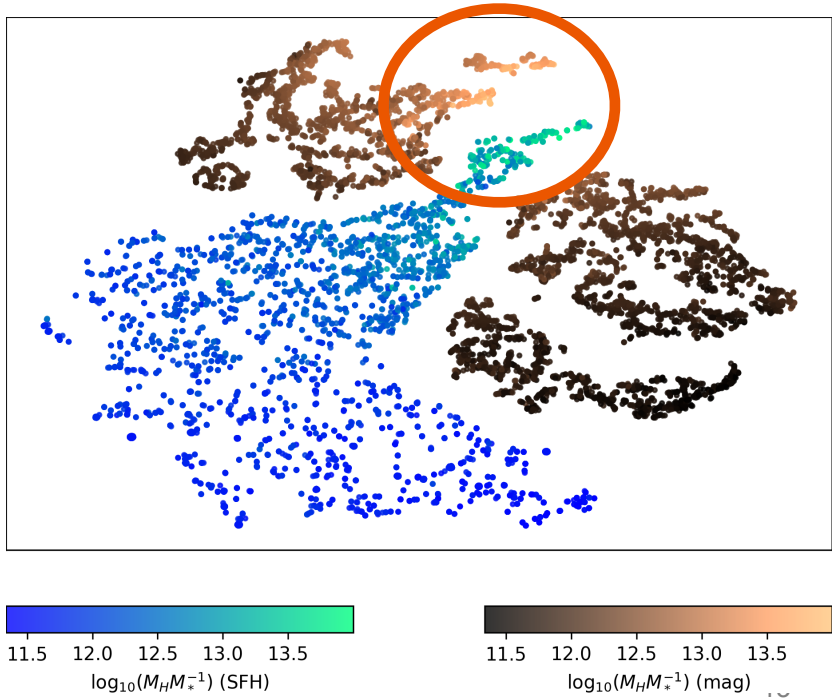
DOMAIN FOUNDATION MODELS



- Training different (encoder-decoder) transformers with different data modalities. Numerical tokenization used.
- Contrastive loss to connect simulation modes with observed/experimental modes
  - Changes embedding space

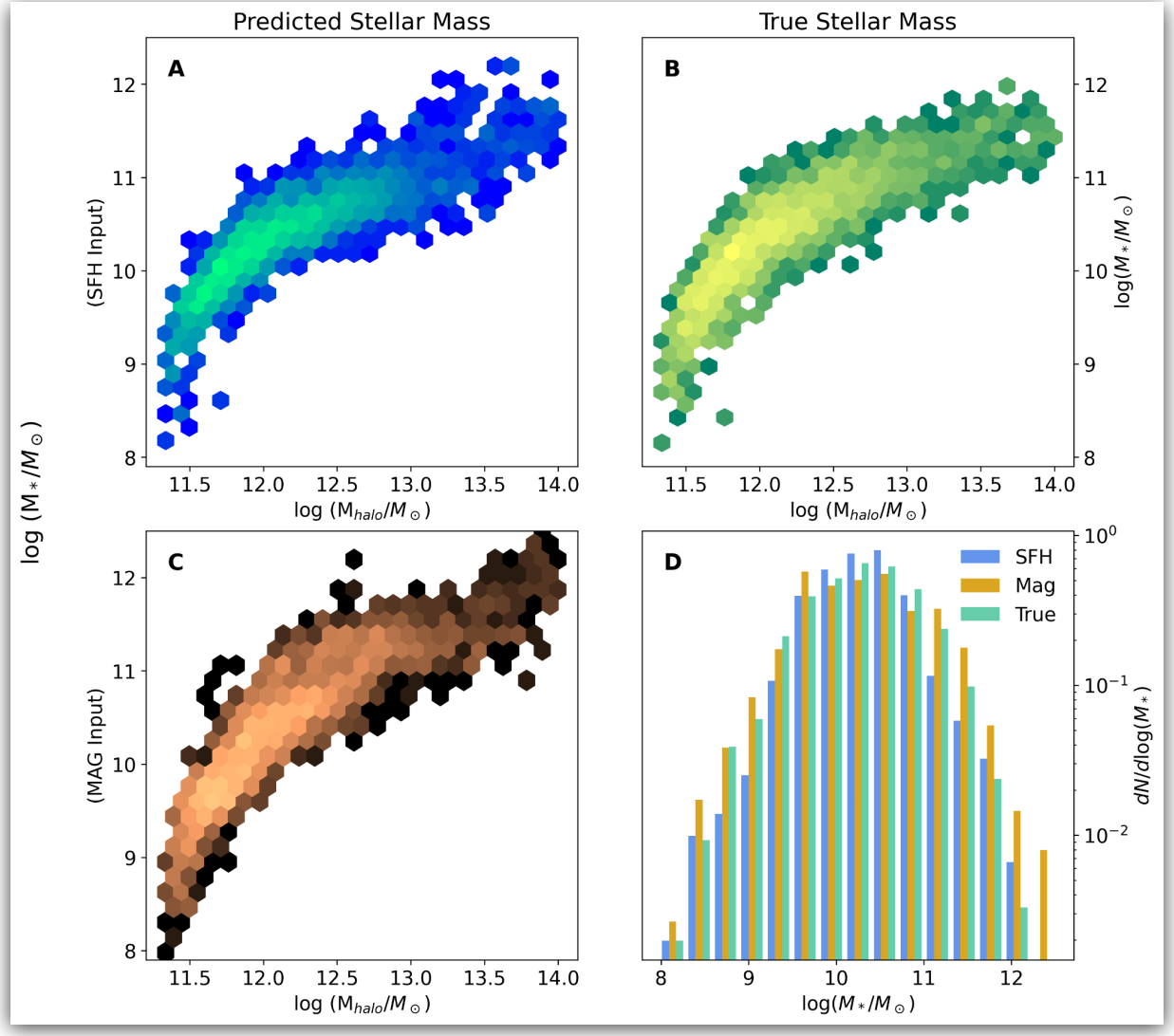


Azton Wells, NR, Salman Habib, in prep



DOMAIN FOUNDATION MODELS

- Training is general purpose, deployment is task-specific.
- Flexibility in deployment: queries dictate latent space access.
- Compatibility wrto datasets in multiple domains
- DFMs can be joined with existing LLMs for contexts along with knowledge base access



Azton Wells, NR, Salman Habib, in prep

CONCLUSIONS AND FUTURE OUTLOOK

- The foundation models have finally facilitated a truly large and deep representations of some of the largest datasets.
- The FMs are mostly trained for language tasks. Encoding non-text information (either directly or indirectly) is the next important step for scientific impact.
- Inputs from science community is crucial for datasets, algorithm development and benchmarking.
- Questions?