AI for Nuclear Physics

PI: Robert Edwards

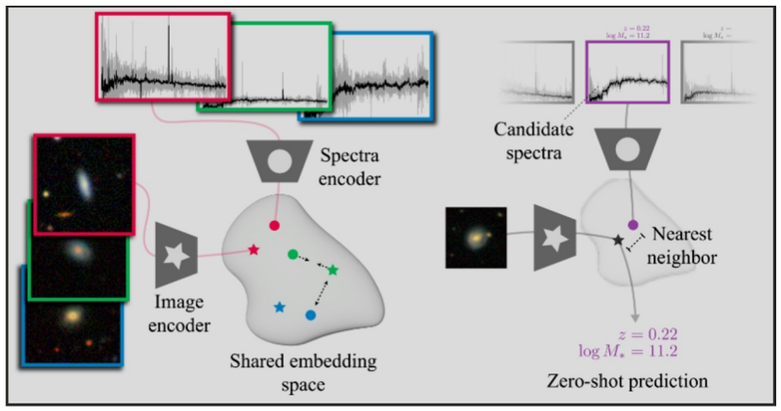
Co-PIs: Diana McSpadden,
Kostas Orginos, Nobuo Sato

Jefferson Lab

Impact and strategic value to the Laboratory's mission

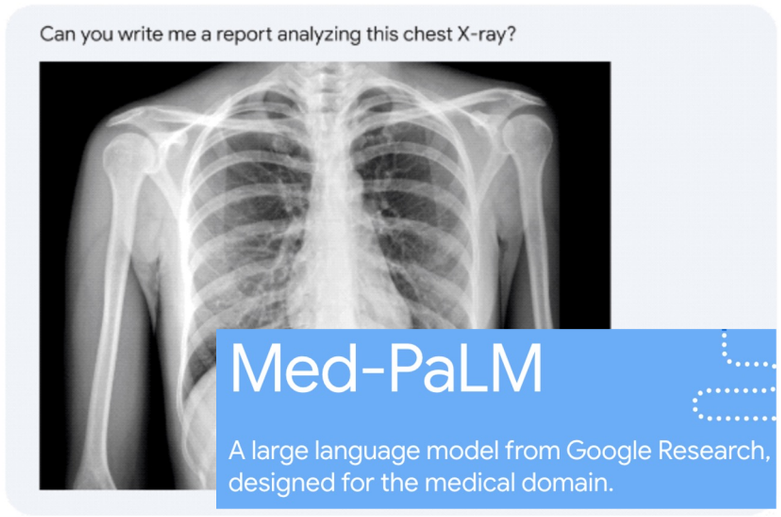
AI & ML having large impact in a diverse array of science

Astrophysics

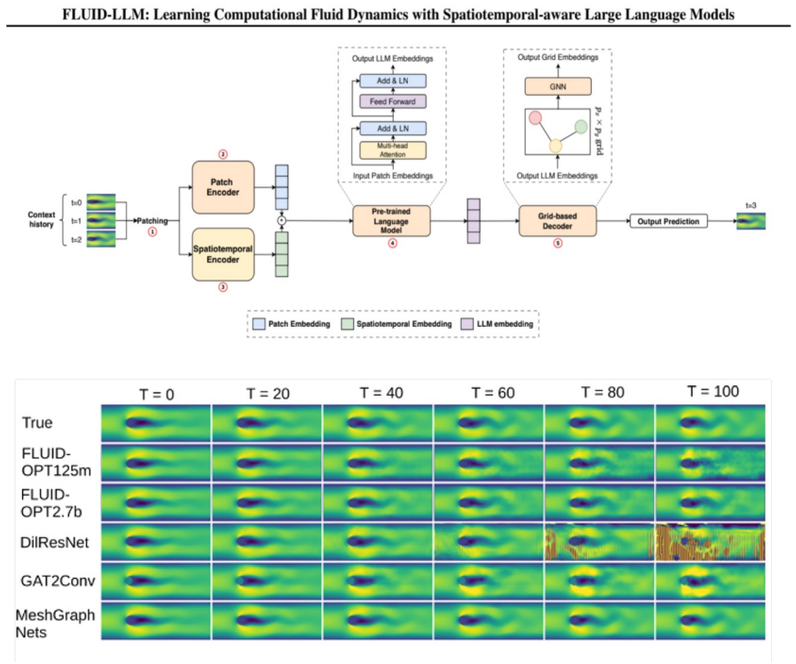


AstroCLIP: single, versatile model that can embed both galaxy images and spectra into a shared, physically meaningful latent space

Medical research



Fluid dynamics



FLUID-LLM: a novel framework combining pre-trained LLMs with spatiotemporal-aware encoding to predict unsteady fluid dynamics

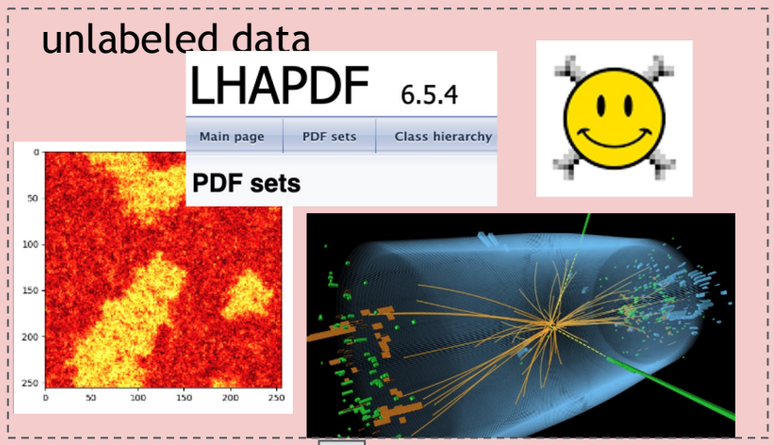
The goal is to build expertise that utilizes artificial intelligence technologies, including Large Language Models, to realize the full potential of JLab domain data and to accelerate the lab's scientific program.

We will establish a new core capability in Foundation Models/LLMs/LLMs for NP

Jefferson Lab

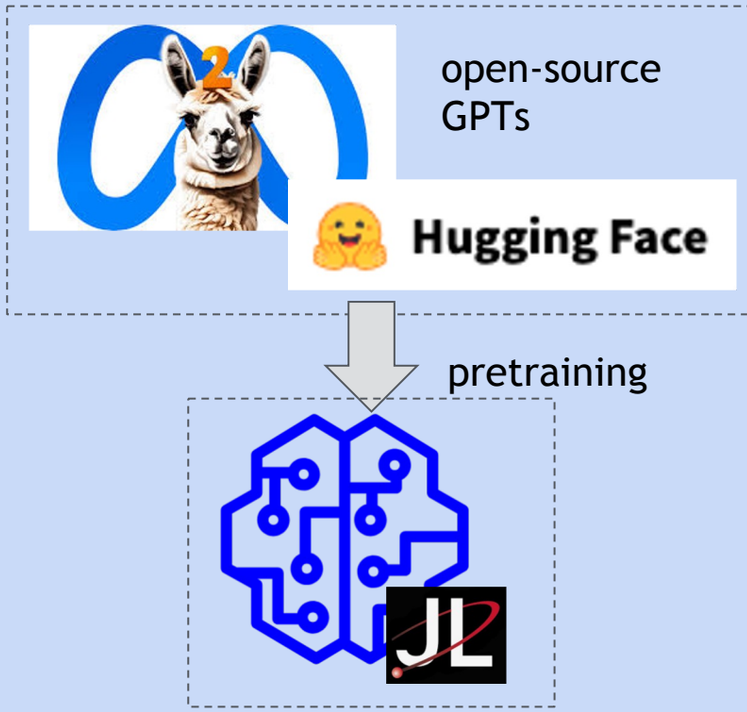
Level of innovation

Aim 1: R&D of data tokenization and embedding models



- Data collection
- Tokenization & simulation DB
- Develop embedding model

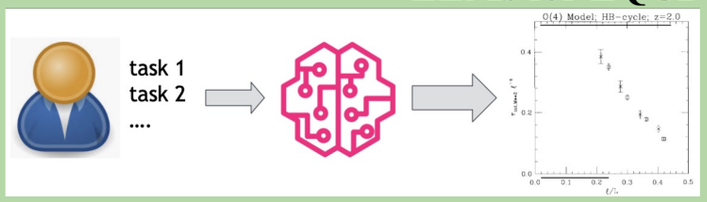
Aim 2: R&D on updating open-source LLMs GPT models, including evaluation



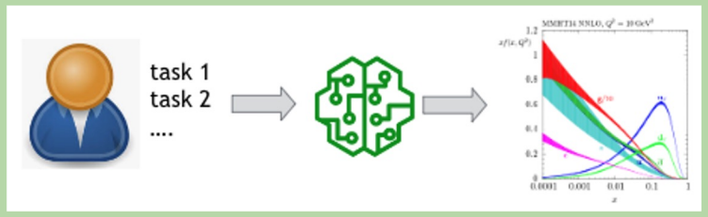
- Additional pretraining of open-source LLMs & evaluation
- Systematic analysis of latent space

Aim 3: R&D on using structured data for the fine tuning of data downstream applications

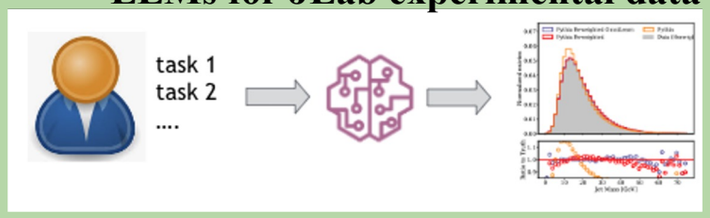
LLMs for LQCD



LLMs for PDFs



LLMs for JLab experimental data



- Task-specific fine-tuning
- Tune LLM & validate

Work cannot be carried out under existing grants - only LDRD

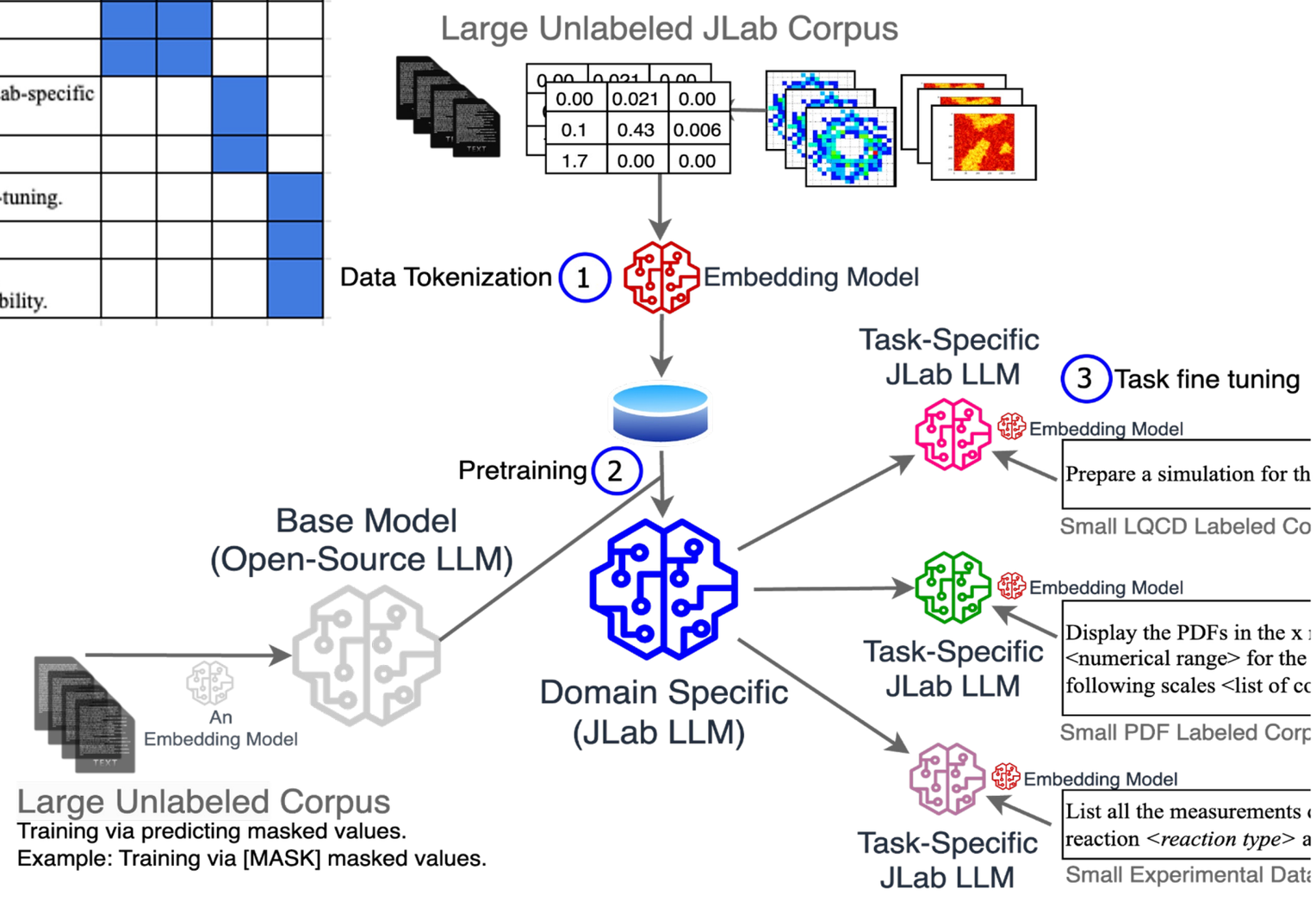
Outcomes:

- Well-placed for future data
- Embed into JLab community
- Disseminate research through seminars/workshops - enable new use cases

Jefferson Lab

Milestones

Aims	Objective number	Milestone	FY24 H1	FY24 H2	FY25 H1	FY25 H2
1	LQCD Milestone 1	Collect observables from spin-model simulations as numeric source for embedding model; Begin formulation of Self-Learning Monte Carlo (SLMC) for the two-dimension O(4) sigma model utilizing Self-Attention transformer; Renormalization Group flow of 2D O(4) nonlinear sigma models via Self-Attention transformers; Utilizing established training codes, characterize potential performance improvements for SLMC compared to established methods				
	PDF Milestone 1	Collect all the data available at LHAPDF; Generate PDF data for all the PDF sets using the LHAPDF interpolation software across a dense grid in x and Q2 along with their corresponding 1-sigma confidence bands; Design suitable tokenization scheme that integrates the info file of the PDF sets, the associated arxiv document and the PDF data; Train and validate the embedding models using the tokenized LHAPDF corpus data.				
	JED Milestone 1	Simulate eP data using Pythia and generate a variety of phase space distributions; Design suitable tokenization scheme that integrates the simulated event level data, the phase space distributions; Train and validate the embedding models; Collect real JLab data along with the associated publications and train and validate the embedding models.				
	Milestone 1	Identify and select research text and numeric data sources from the use cases				
	Milestone 2	The tokenized text and simulation data database from Milestone 1				
	Milestone 3	Develop a prototype of the embedding model.				
2	Milestone 1	Train the base, open-source LLMs in an unsupervised manner using JLab-specific unlabeled data, specifically, the vector database developed in Aim 1.				
	Milestone 2	Carry out a systematic analysis of the latent space with existing tools.				
3	Milestone 1	For the three use cases, develop the questions for the task-specific fine-tuning.				
	Milestone 2	Tune the LLM and validate the data from Milestone 1.				
	Milestone 3	Evaluate the trained LLM from initial LLM and post-transfer learning. Evaluations will include accuracy, biases, training time, and generalizability.				



Outcomes:

- Well-placed for future data
- Embed into JLab community
- Disseminate research through seminars/workshops - encourage new use cases

Jefferson Lab

Project team & responsibilities

All Team members contribute to the common milestones of Aims 1, 2, 3

Theory Group:

Robert Edwards: Will oversee the project. Apply LLM methods to constructing improved simulation methods in LQCD use case of Aim 1.

Kostas Orginos: AI/ML Simulation methods in Aim 1 for LQCD use case.

Nobuo Sato: Data encoding methods - connecting theory to expt. for PDF and JED use-case for year 1.

Data Group:

Diana McSpadden: Characterization of data encoding for all use cases. Will contribute to common year 1 milestones, and all year 2 milestones.

Postdoc:

TBD : Data and Theory Simulation analyst. Contribute to data-encoding of PDF & JLab Expt Data (JED) use cases. This 1.0 FTE could support two people.

Jefferson Lab

Budget & justification

Different components under Aim 1 (1st year)

All team members share common milestones for Aims 2 & 3 (2nd year)

Team member	Role	Year 1	Year 2	Contribution
Robert Edwards	PI	20%	20%	Aim 1: LLM methods to construct improved simulation methods for LQCD
Diana McSpadden	Co-PI	40%	40%	Characterization of data encoding. Lead pre-training of LLMs & evaluation
Kostas Orginos	Co-PI	20%	0	AI/ML simulations in Aim 1. Common milestones in year 2
Nobuo Sato	Co-PI	5%	5%	Data encoding methods: theory-expt for PDF & Events use-case in Aim 1
TBD	Postdoc	100%	100%	Data & Theory simulation analyst

Budget ($K)	Total	FY24	FY25	FY26
	562	0	294	268

Jefferson Lab

Potential future funding (Beyond LDRD)

FOA:

- DOE/NP has plans to release in the fall an FOA on the topic *"AI in Nuclear Physics"*
- As understood, intention is for NP to build capability & utilize AI for range of topics
- Likely LLMs will be central to the call

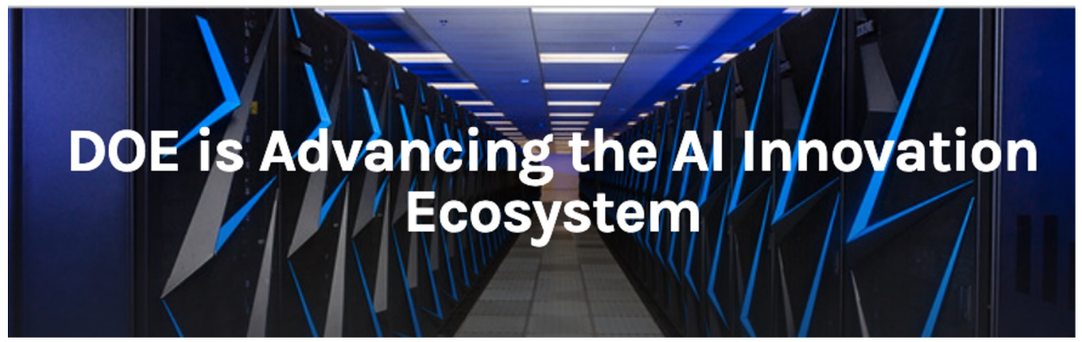
Potential future directions for proposals:

- Application of Attention technologies for LGT simulations
- R&D for Foundational Models that integrate coherently all data from JLab for Expt.& Theory
- Development of AI Agents for Nuc. Theory research

Why us? Why now?

- Lab needs experience & capabilities
- AI or Die…

Office of
CRITICAL AND EMERGING TECHNOLOGIES



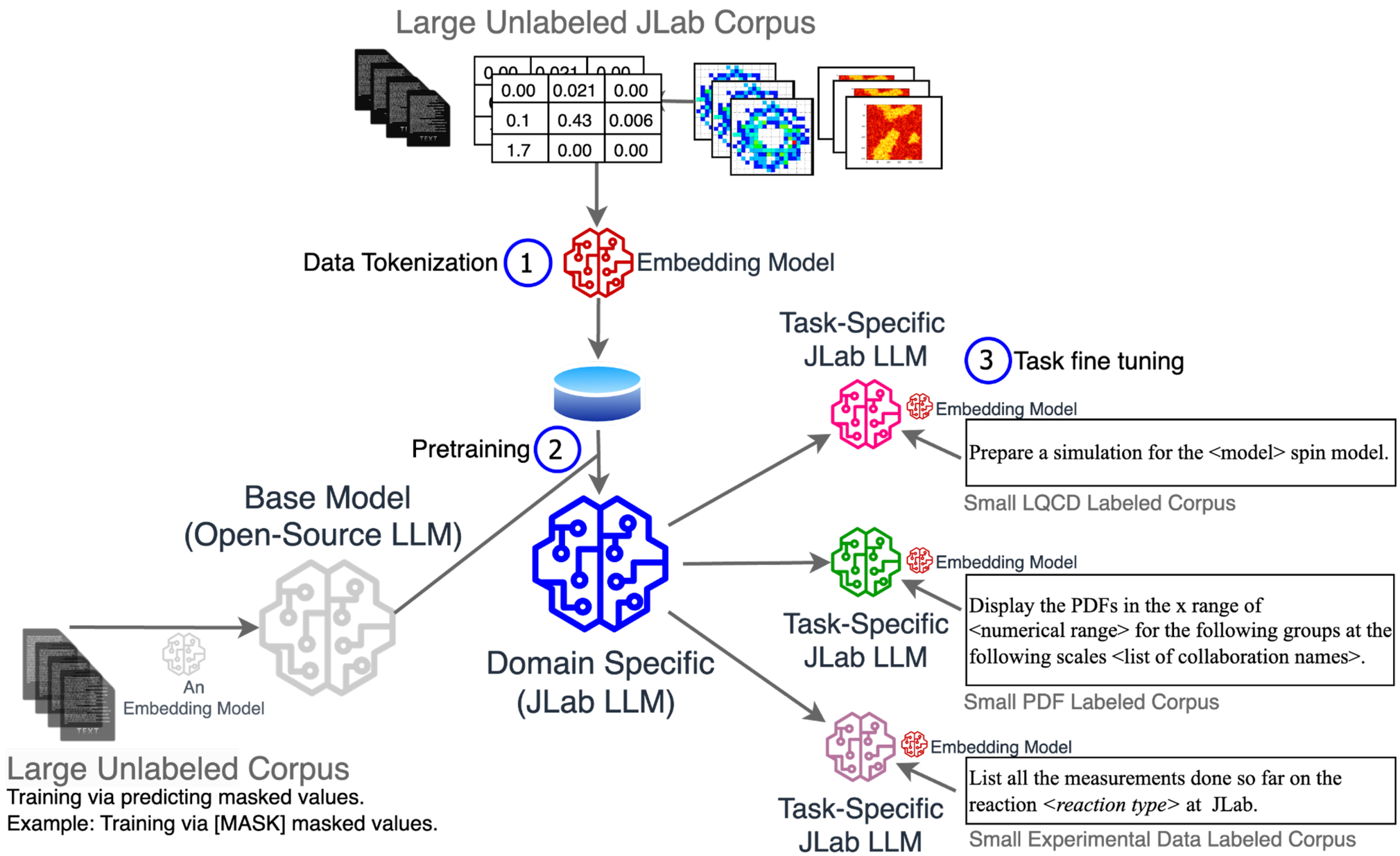
DOE Press Releases

- U.S. Department of Energy Announces $18 Million to Advance Particle Accelerator Technologies and Workforce Training (March 2021)
- DOE Announces $37 Million for Artificial Intelligence and Machine Learning at DOE Scientific User Facilities (August 2020)

NP Funding Opportunity Announcements & Awards Lists

- Data, Artificial Intelligence, and Machine Learning at Scientific User Facilities (LAB 20-2261 FOA), (Award List), 2020.
- Data Analytics for Autonomous Optimization and Control of Accelerators and Detectors (FOA-0002490), 2021.
- Artificial intelligence and Machine Learning for Autonomous Optimization and Control of Accelerators and Detectors (DE-FOA-0002875), 2023.

Organization



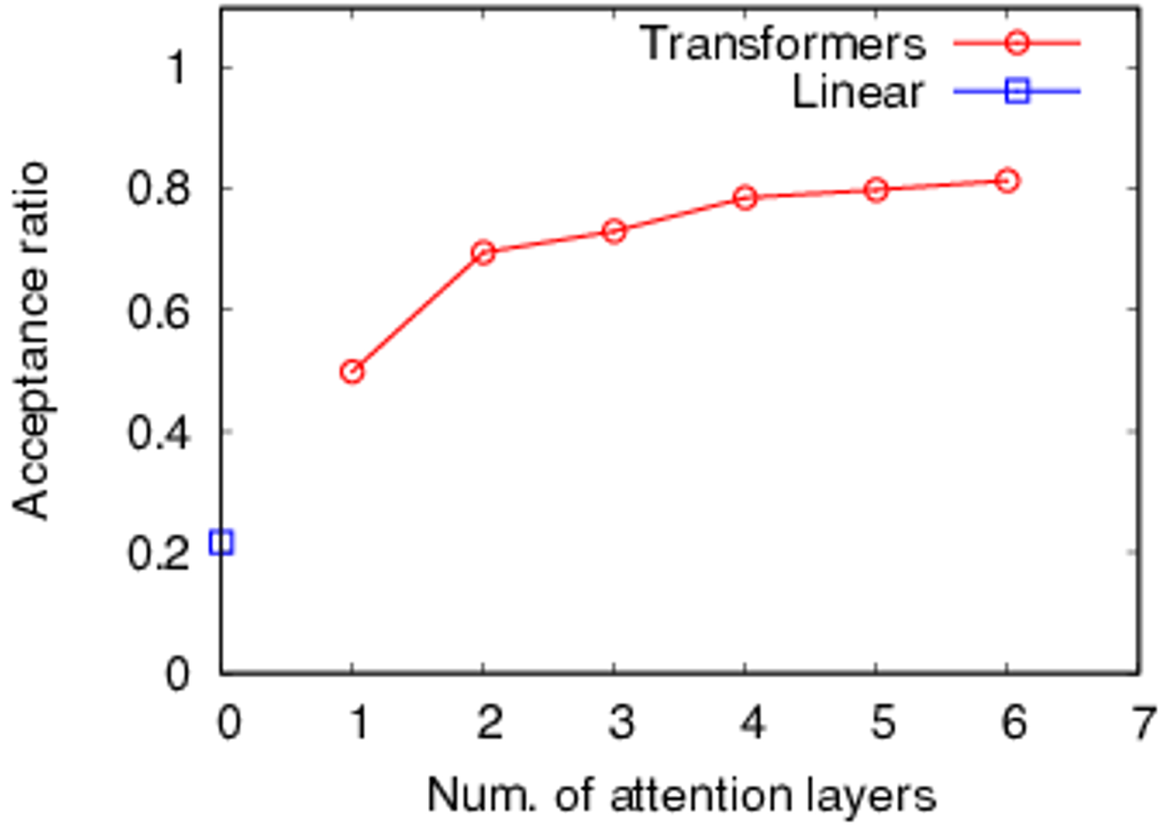
Example: Self-Attention Transformers

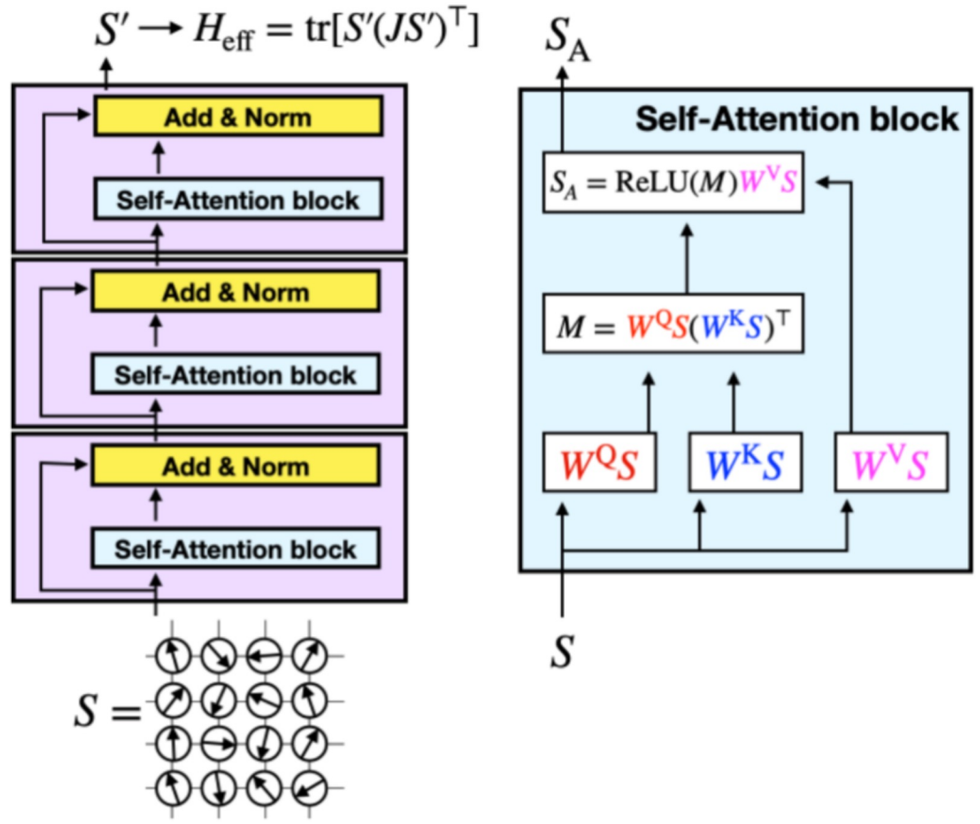
Accessing smaller lattice spacings critical to physics program.
Critical slowing down - exponential growth in cost.
Eliminated in a 2d fermionic model.

Attention: https://doi.org/10.48550/arXiv.1706.03762

Self-learning Monte Carlo with equivariant Transformer

https://arxiv.org/pdf/2306.11527

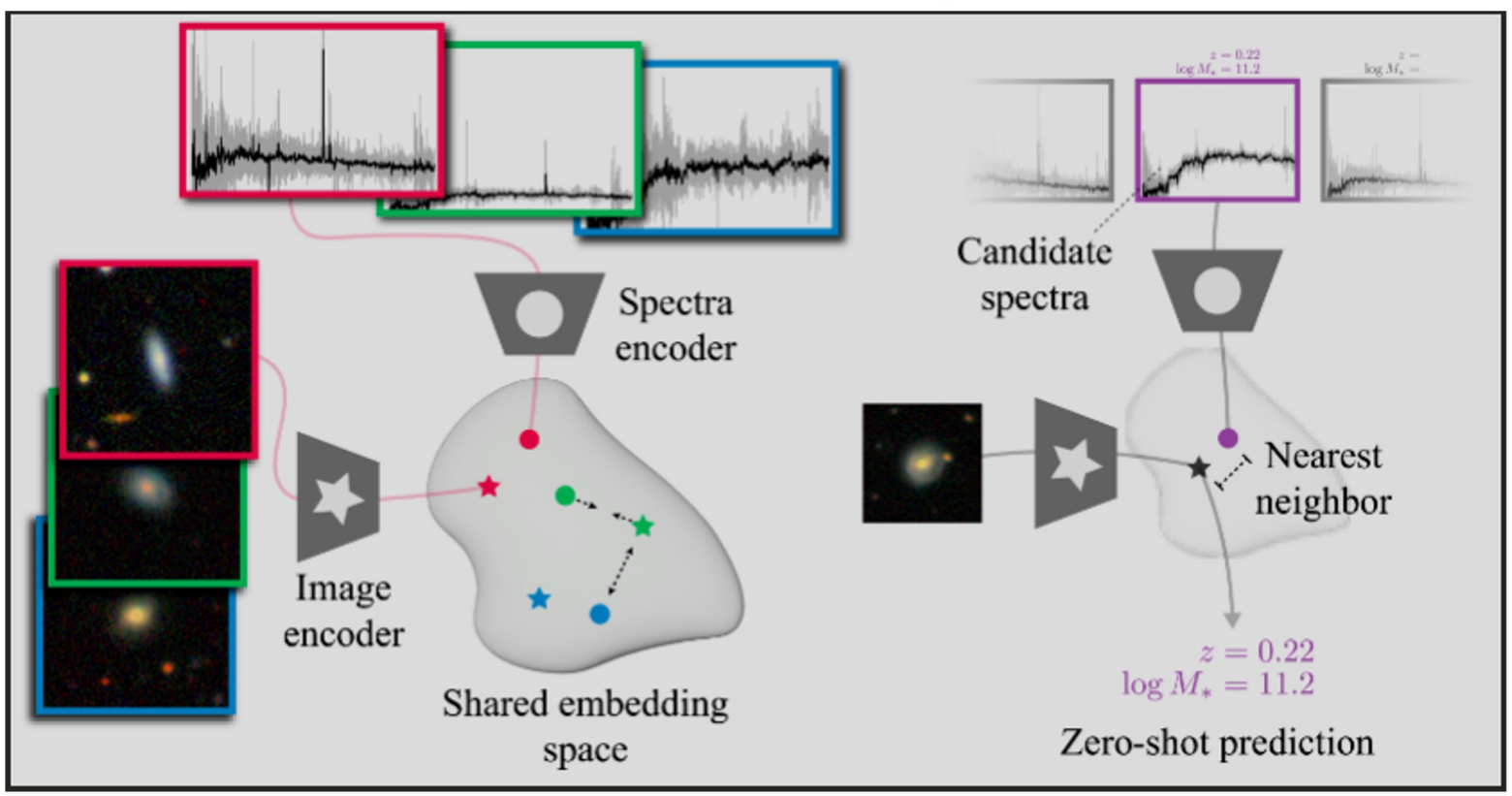




Jefferson Lab

Example: AstroCLIP

AstroCLIP https://doi.org/10.1093/mnras/stae1450



https://arxiv.org/html/2310.03024v2

Figure 1: Illustration of the *AstroCLIP* cross-modal training strategy. This approach consists of two steps. First, galaxy images and spectra are embedded separately by pretraining both an image and a spectrum encoder in a SSL setting. Then, these encoders are aligned using a cross-modal contrastive loss. Once aligned, these embeddings allow us to connect and compare cross-modal representations. Moreover, they possess physically meaningful high-level information which can be used for a variety of downstream tasks on which the model was neither trained nor fine-tuned.

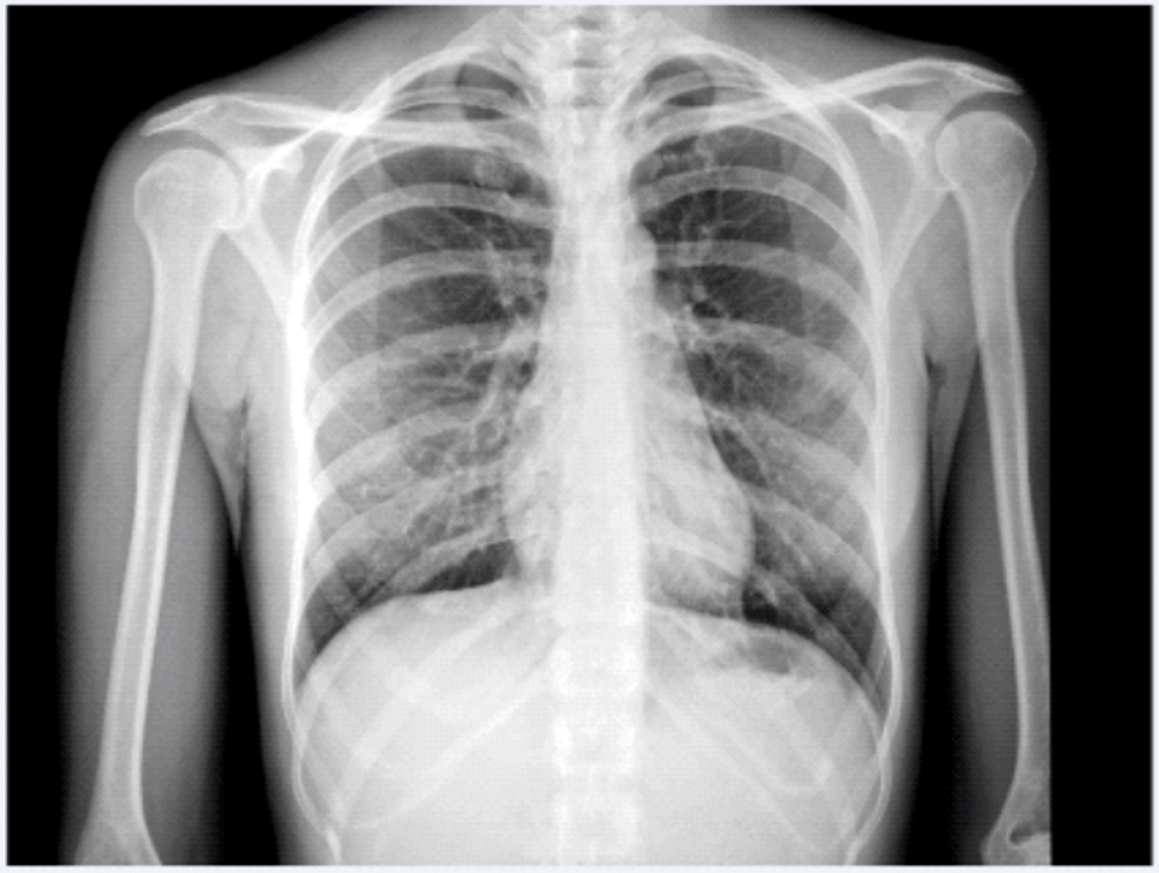
Example: Med-PaLM

Med-PaLM

A large language model from Google Research, designed for the medical domain.

Singhal, K., Azizi, S., Tu, T. *et al.* Large language models encode clinical knowledge. *Nature* 620, 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2

Can you write me a report analyzing this chest X-ray?



Findings:

- Devices: None.
- Lungs: No pneumothorax. No substantial pleural effusion. Lungs appear clear.
- Cardiomediastinal: Normal heart size. Mediastinal contours within normal limits.
- Other: No acute skeletal abnormality.

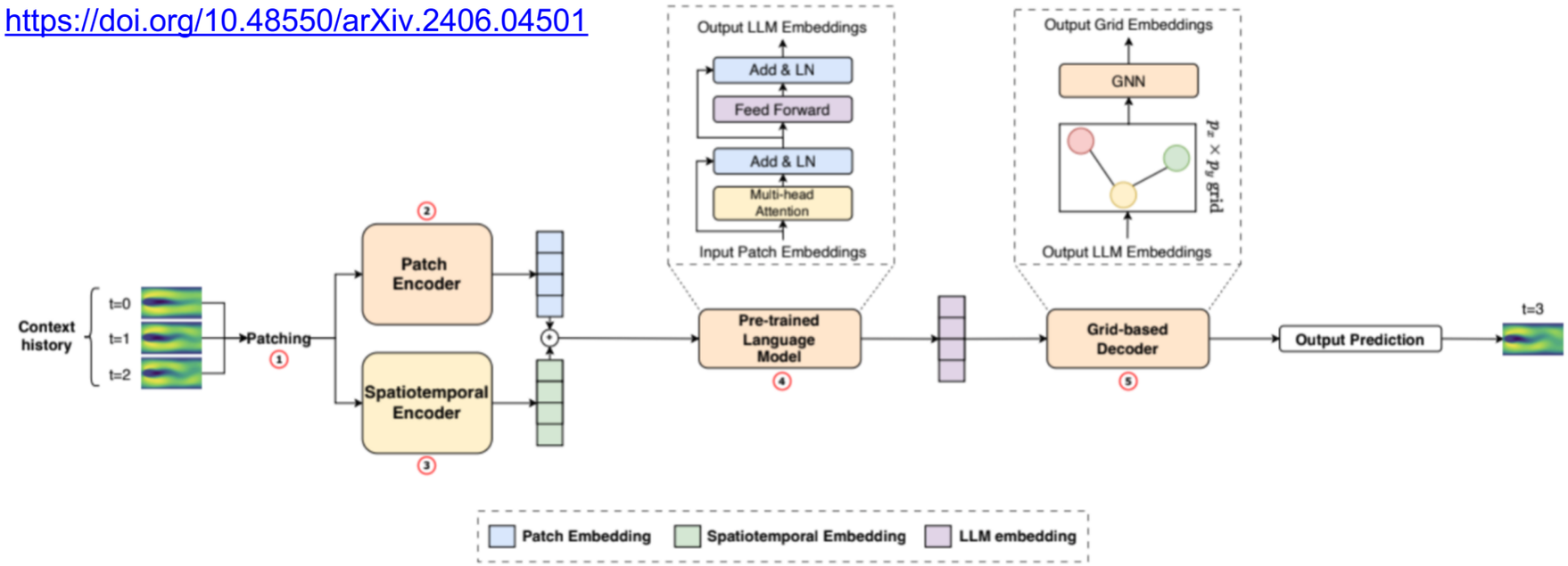
Impression:

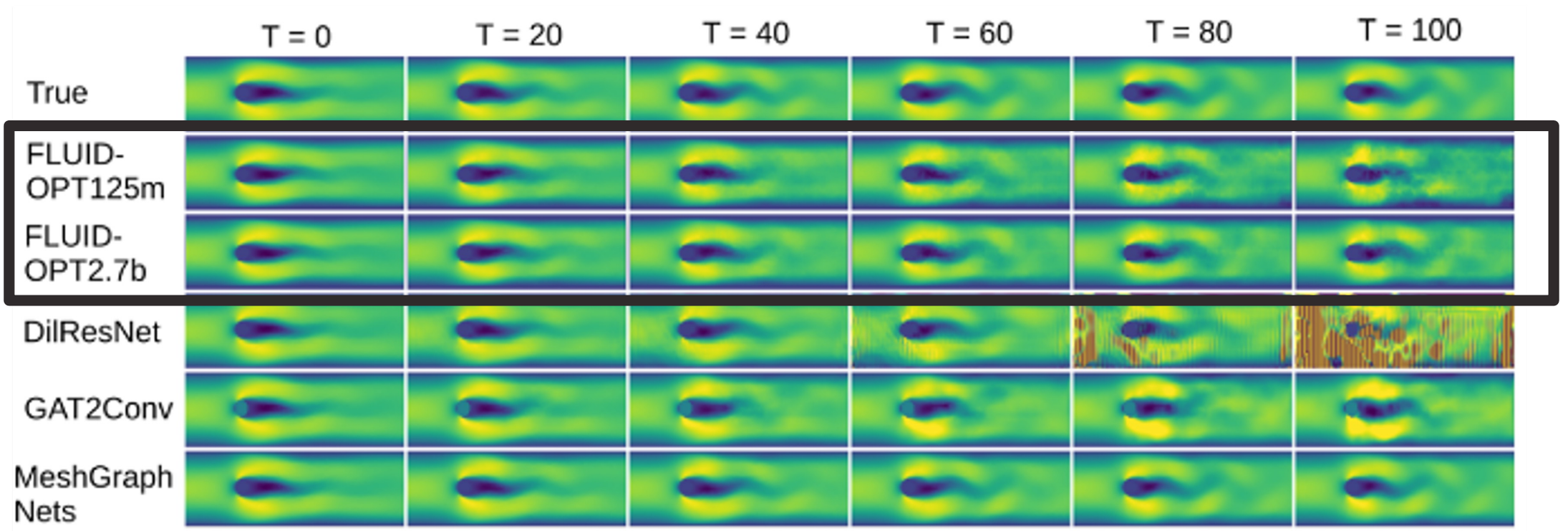
No active disease seen in chest.

Example: FluidLLM

FLUID-LLM: Learning Computational Fluid Dynamics with Spatiotemporal-aware Large Language Models

https://doi.org/10.48550/arXiv.2406.04501





Jefferson Lab

Info on presentations

The presentations of the FY25 LDRD proposals are scheduled for Friday, July 26, 2024, starting at 11:00 AM. The agenda is attached. Each presentation is scheduled for exactly 10 minutes, followed by 5 minutes of questions from the committee. Susan is creating an Indico page where you will need to upload your presentation by July 26 at 10:00 AM.

In addition to the project description, scope, and team, your presentations should address the following:

- Impact and strategic value to the Laboratory’s mission
- Level of innovation
- Deliverables with corresponding timeline and milestones
- Budget and budget justification
- Potential future funding (Beyond LDRD)

The meeting will be in B207, and you will need to be there only for your presentation. Let me know if you have any questions.