VILLUM FONDEN

Co-financed by the Connecting Europe Facility of the European Union

INTERACTIONS

PolarBERT: A Foundation Model for IceCube

UNIVERSITY OF COPENHAGEN

Inar Timiryasov, Jean-Loup Tastet, Oleg Ruchayskiy

Niels Bohr Institute and DIKU, University of Copenhagen

Nikhef Theory Seminar

2025-06-25, Amsterdam

Plan

- IceCube introduction
- Foundation models
- PolarBert (foundation model for IceCube)
- But what is a transformer model?
- LLMs and physics (a very opinionated part)

* I am not a member of IceCube

My background

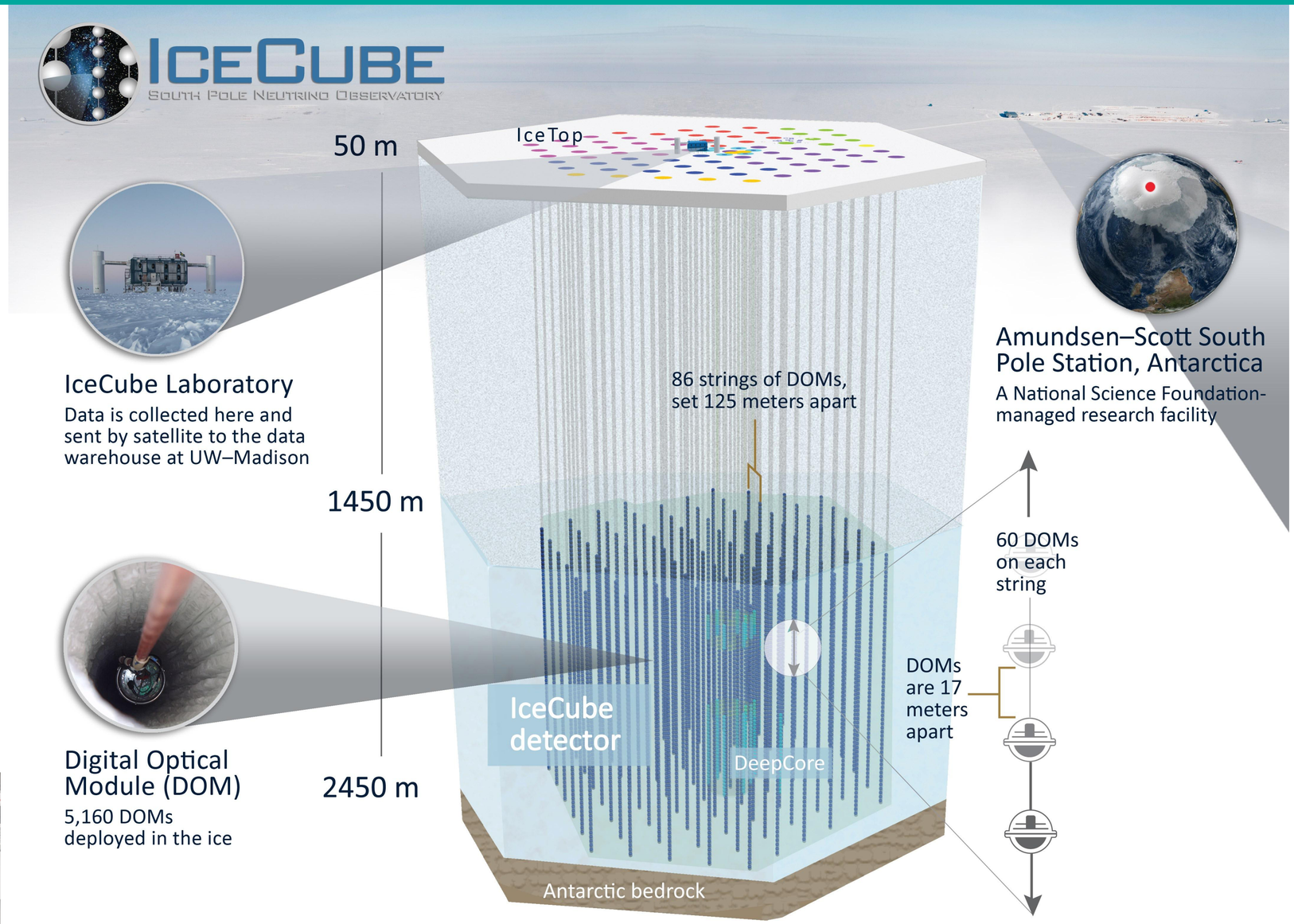
- Models with large extra dimensions
- SHiP phenomenology
- Leptogenesis
- Einstein-Cartan gravity (inflation and dark matter)
- Machine Learning (Continuous gravitational waves and IceCube Kaggle competitions)
- Language Models (arxiv.org/abs/2308.02019, arxiv.org/abs/2409.17312)
- PolarBert (a foundation model for IceCube, paper link)

* I am not a member of IceCube

IceCube

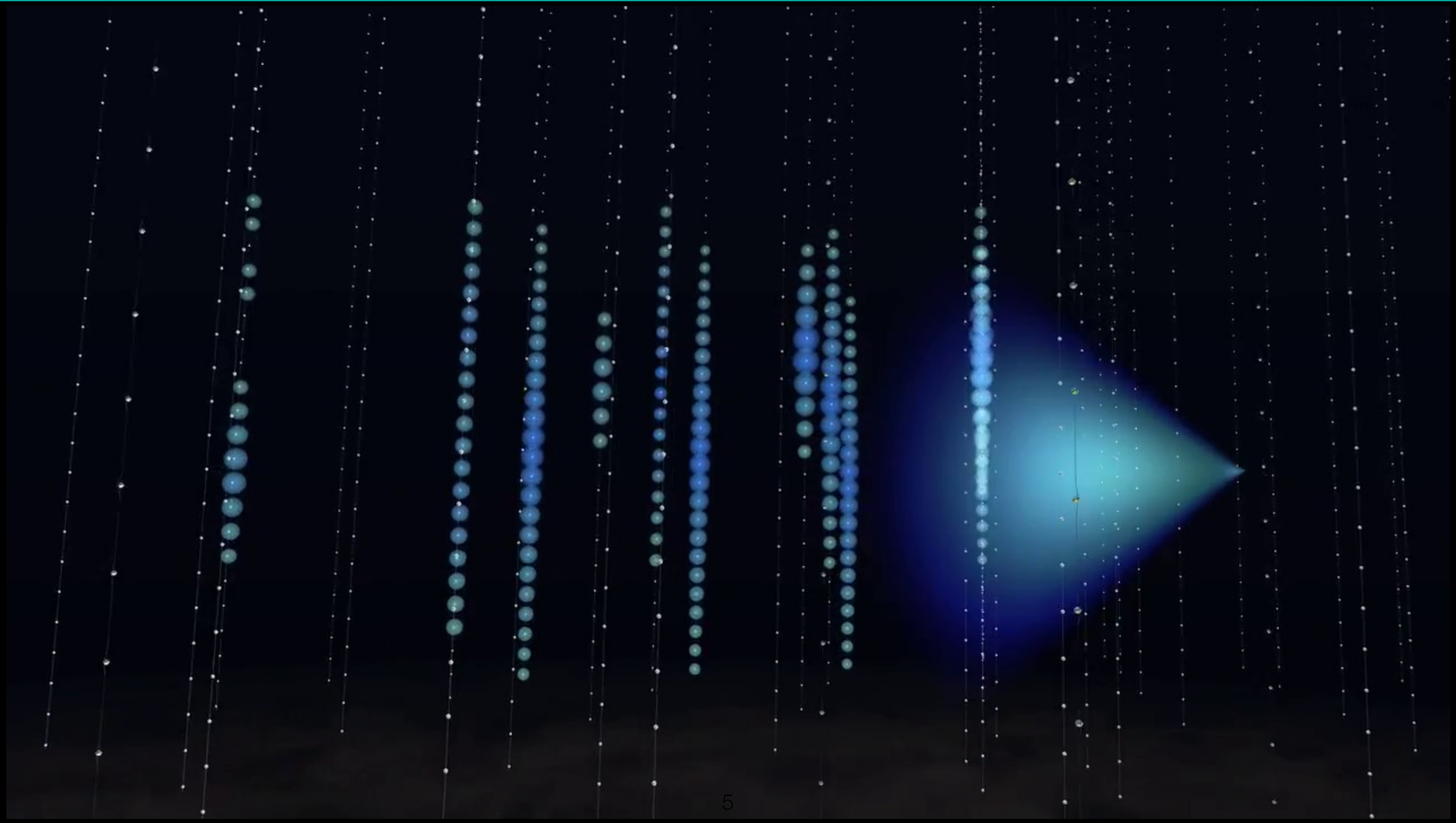
- Neutrino telescope
- Located at the South Pole
- Detector volume: 1 cubic kilometer
- Oftentimes observes through Earth
- 5160 optical modules (DOMs)
- Public dataset from Kaggle Competition 130 million events

KM3NeT module:



IceCube event

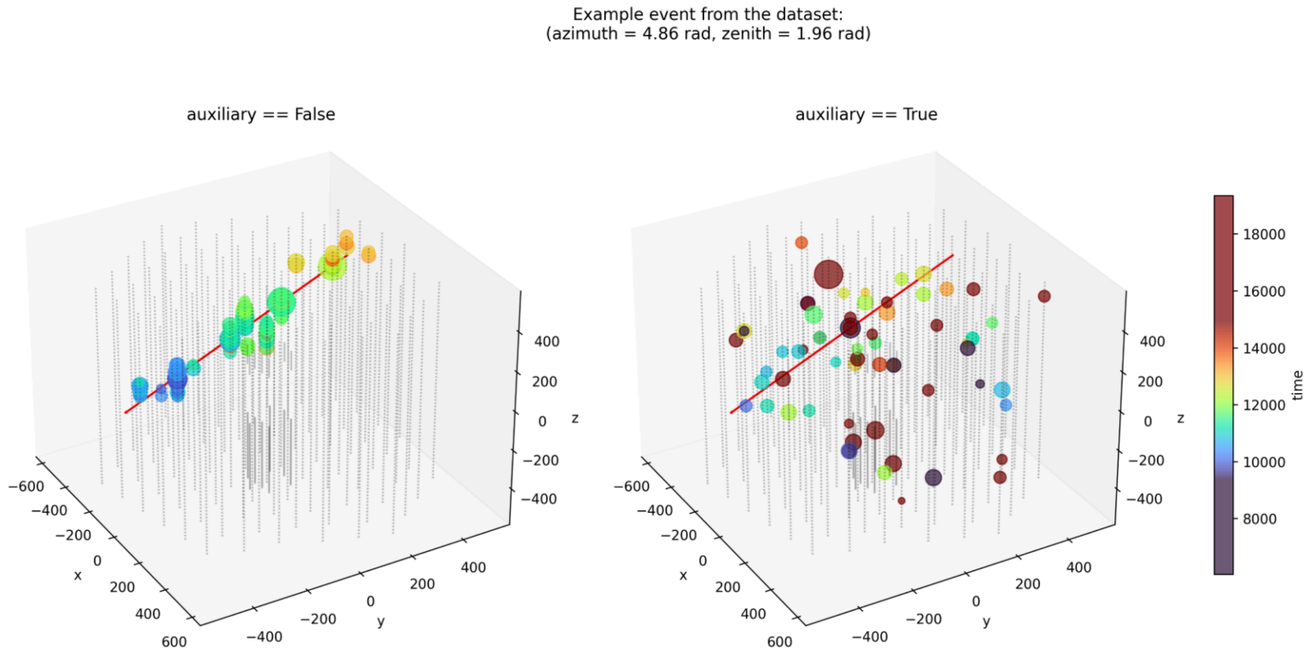
https://youtu.be/OXSqiPLn9CM?si=nnvKH0WpJgEWRn56



Inverse problem: reconstruct the neutrino direction

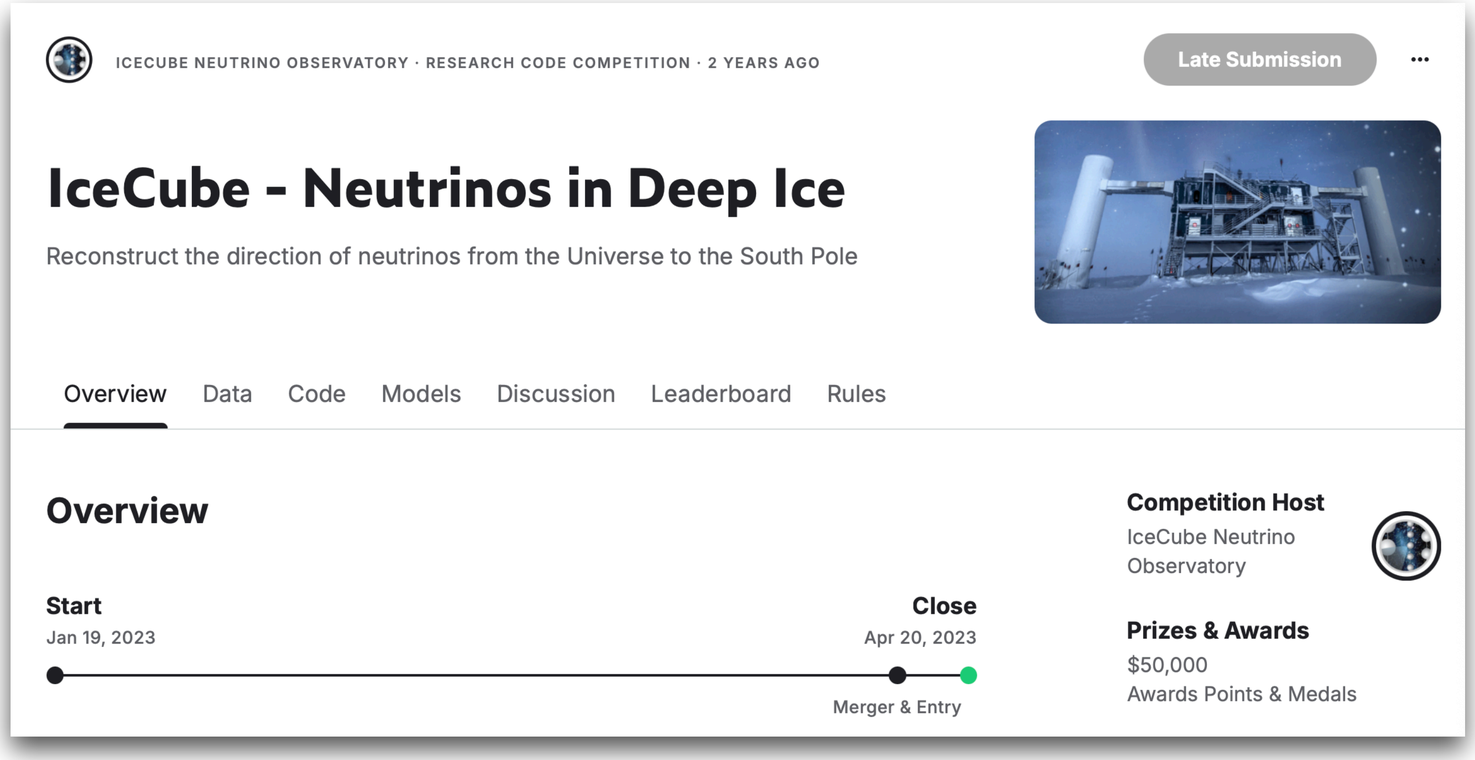
- Neutrino energy
- Neutrino direction
  (astrophysical sources; identification with galactic plane)
- Traditional methods: likelihood based
- $L(x, y, z, t, \theta, \phi) = p(\text{data} \,|\, x, y, z, t, \theta, \phi)$
- $$L(x, y, z, t, \theta, \phi) = \prod_{j=1}^{N_{DOM}} \prod_{i=1}^{N_{hit}} [p_j(t_i)]^{q_i}$$

  t_i - pulse time, q_i - charge
- To maximize the likelihood one has to simulate light propagation through Ice
  (currently used: arxiv.org/abs/2103.16931)



Machine Learning in IceCube

- Graph Neural Networks for Low-Energy Event Classification & Reconstruction in IceCube
  https://arxiv.org/abs/2209.03042
- A Kaggle competition in 2023
  (901 Participants)
- Kaggle is a specialized platform
  for ML competitions
- Still not better than
  traditional methods at high energies



Can we learn something from LLM progress?

- LLMs benefit from internet-scale datasets.
- Physics also has a lot of data.
  - Both labeled (MC) and unlabeled.
- Can we benefit from unlabeled data?

An exabyte of disk storage at CERN

CERN disk storage capacity passes the threshold of one million terabytes of disk space

29 SEPTEMBER, 2023 | By Tim Smith

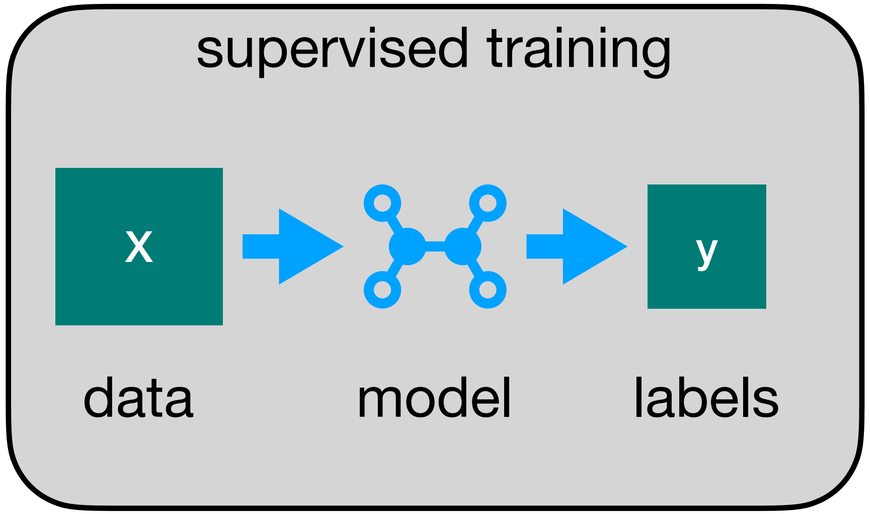
A fraction of the 111 000 devices that form CERN's data storage capacity. (Image: CERN)

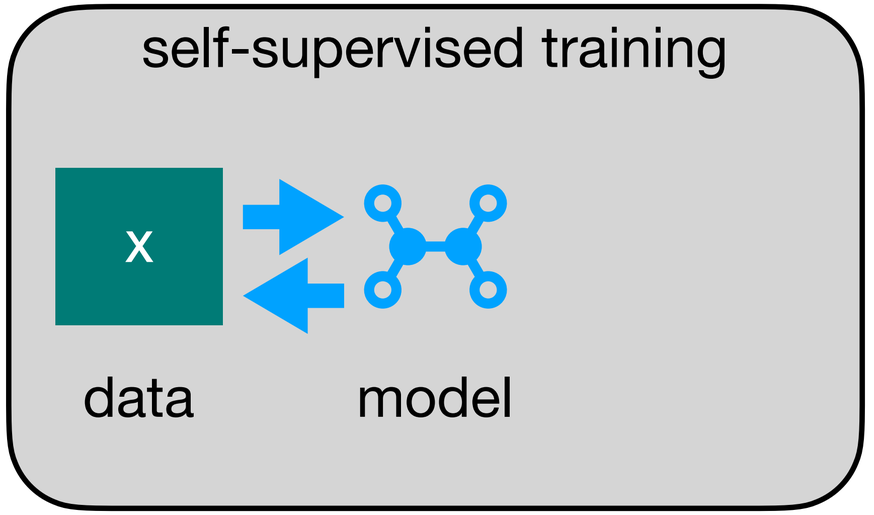
source:
https://home.cern/news/news/computing/exabyte-disk-storage-cern

What do we mean by “foundation models”?

- Initially, the term has been coined for models like BERT and GPT-3
  2108.07258 “On the Opportunities and Risks of Foundation Models”

- Here, by foundational models we mean the models that are pretrained in a self-supervised way and can be fine-tuned for downstream tasks.

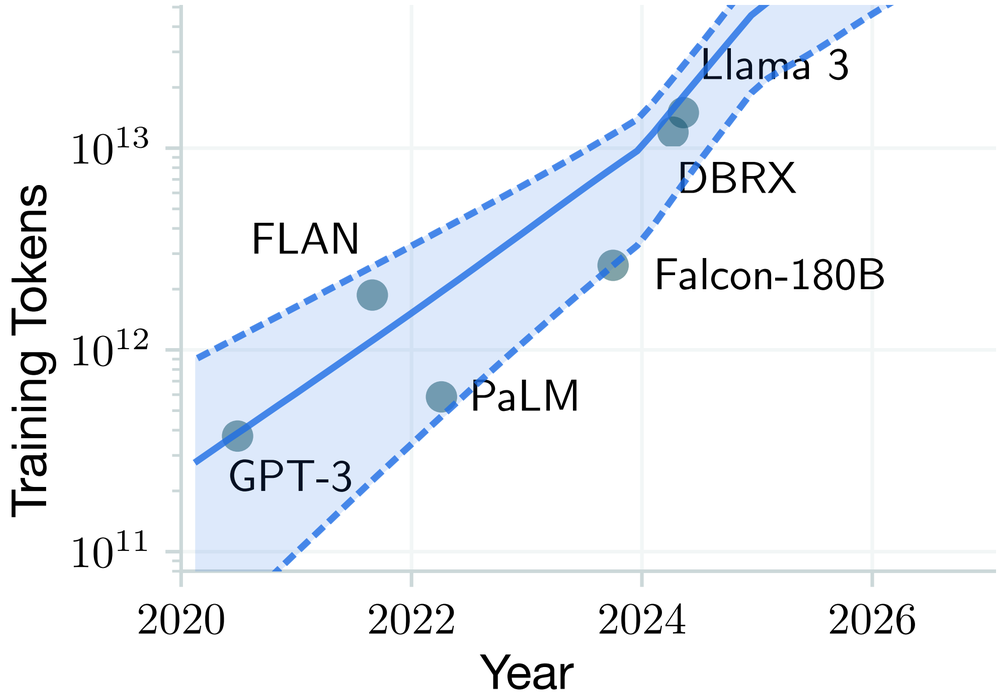




Success of self-supervise training

Outside physics:

- Labeled data is limited
- Unlabeled data is abundant (text, image, video)
- Led to GenAI revolution



source:
2211.04325 "Will we run out of data?
Limits of LLM scaling based on human-generated data"

● BERT - 3.3B tokens

1810.04805 "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"

Success of self-supervise training

- No signs of stopping!

Power

Probably the single biggest constraint on the supply-side will be power. Already, at nearer-term scales (1GW/2026 and especially 10GW/2028), power has become the binding constraint: there simply isn't much spare capacity, and power contracts are usually long-term locked-in. And building, say, a new gigawatt-class nuclear power plant takes a decade. (I'll wonder when we'll start seeing things like tech companies buying aluminum smelting companies for their gigawatt-class power contracts.[57])

https://situational-awareness.ai/
Leopold Aschenbrenner, June 2024

MICROSOFT / TECH / SCIENCE

Microsoft wants Three Mile Island to fuel its AI power needs

Photo by Andrew Caballero-Reynolds / AFP via Getty Images

/ Microsoft has signed a 20-year deal to exclusively access 835 megawatts of energy from a nuclear plant.

By Tom Warren, a senior editor and author of *Notepad*, who has been covering all things Microsoft, PC, and tech for over 20 years.

Sep 20, 2024 at 2:23 PM GMT+2

69 Comments (69 New)

https://www.theverge.com/2024/9/20/24249770/

Self-supervise training: Scaling Laws

Performance predictably improves with scale

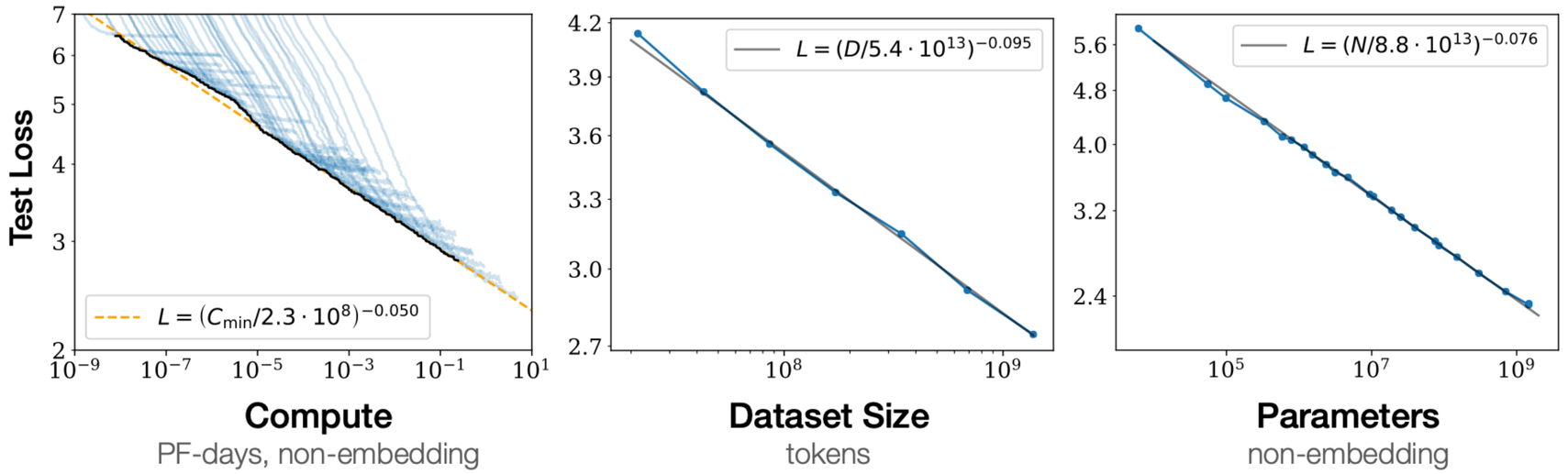


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

https://arxiv.org/pdf/2001.08361
Scaling Laws for Neural Language Models
Jared Kaplan et al

Foundation models in particle physics

(a very incomplete list)

- **Pre-training strategy using real particle collision data for event classification in collider physics**
  https://arxiv.org/abs/2312.06909
  Tomoe Kishimoto, Masahiro Morinaga, Masahiko Saito, Junichi Tanaka

- **Finetuning Foundation Models for Joint Analysis Optimization**
  https://arxiv.org/abs/2401.13536
  Matthias Vigl, Nicole Hartman, Lukas Heinrich

- **Masked Particle Modeling on Sets: Towards Self-Supervised High Energy Physics Foundation Models**
  https://arxiv.org/abs/2401.13537
  Lukas Heinrich, Tobias Golling, Michael Kagan, Samuel Klein, Matthew Leigh, Margarita Osadchy, John Andrew Raine

- **A Language Model for Particle Tracking**
  https://arxiv.org/abs/2402.10239
  Andris Huang, Yash Melkani, Paolo Calafiura, Alina Lazar, Daniel Thomas Murnane, Minh-Tuan Pham, Xiangyang Ju

- **OmniJet-α: The first cross-task foundation model for particle physics**
  https://arxiv.org/abs/2403.05618
  Joschka Birk, Anna Hallin, Gregor Kasieczka

- **Re-Simulation-based Self-Supervised Learning for Pre-Training Foundation Models**
  https://arxiv.org/abs/2403.07066
  Philip Harris, Michael Kagan, Jeffrey Krupa, Benedikt Maier, Nathaniel Woodward

- **OmniLearn: A Method to Simultaneously Facilitate All Jet Physics Tasks**
  https://arxiv.org/abs/2404.16091
  Vinicius Mikuni, Benjamin Nachman

Physics > Data Analysis, Statistics and Probability

[Submitted on 9 Jan 2025]

Large Physics Models: Towards a collaborative approach with Large Language Models and Foundation Models

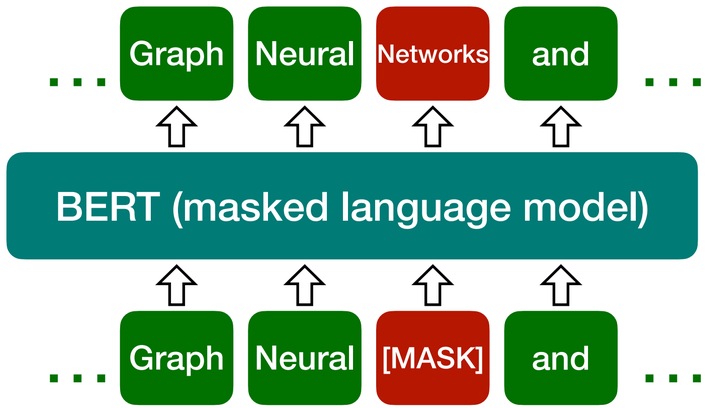
Kristian G. Barman, Sascha Caron, Emily Sullivan, Henk W. de Regt, Roberto Ruiz de Austri, Mieke Boon, Michael Färber, Stefan Fröse, Faegheh Hasibi, Andreas Ipp, Rukshak Kapoor, Gregor Kasieczka, Daniel Kostić, Michael Krämer, Tobias Golling, Luis G. Lopez, Jesus Marco, Sydney Otten, Pawel Pawlowski, Pietro Vischia, Erik Weber, Christoph Weniger

https://arxiv.org/abs/2501.05382

Challenges of self-supervise learning in particle physics

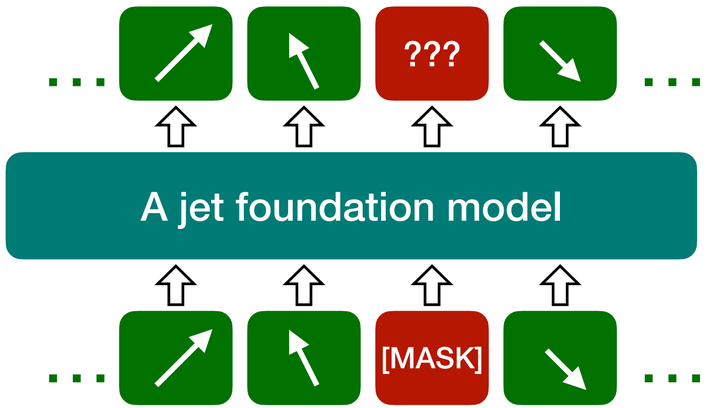
BERT
(Bidirectional Encoder Representations from Transformers)

predict the distribution of a token from a discrete set



A jet foundation model

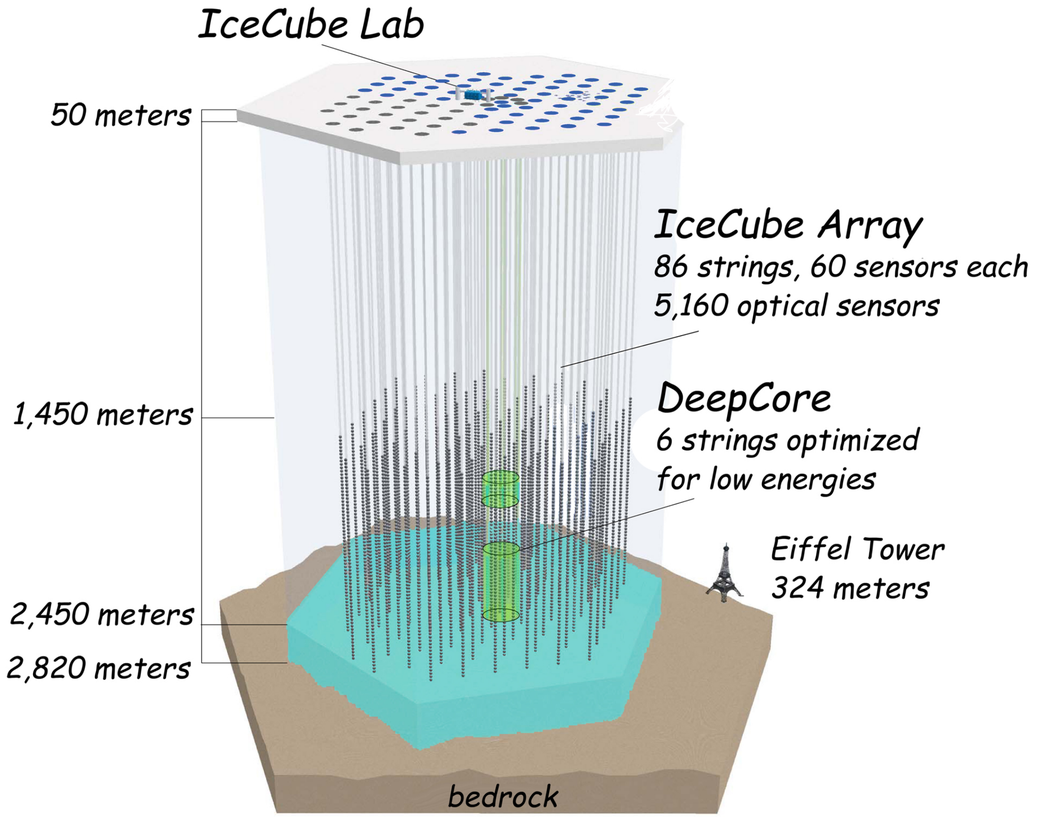
How to predict a continuous 4-vector?



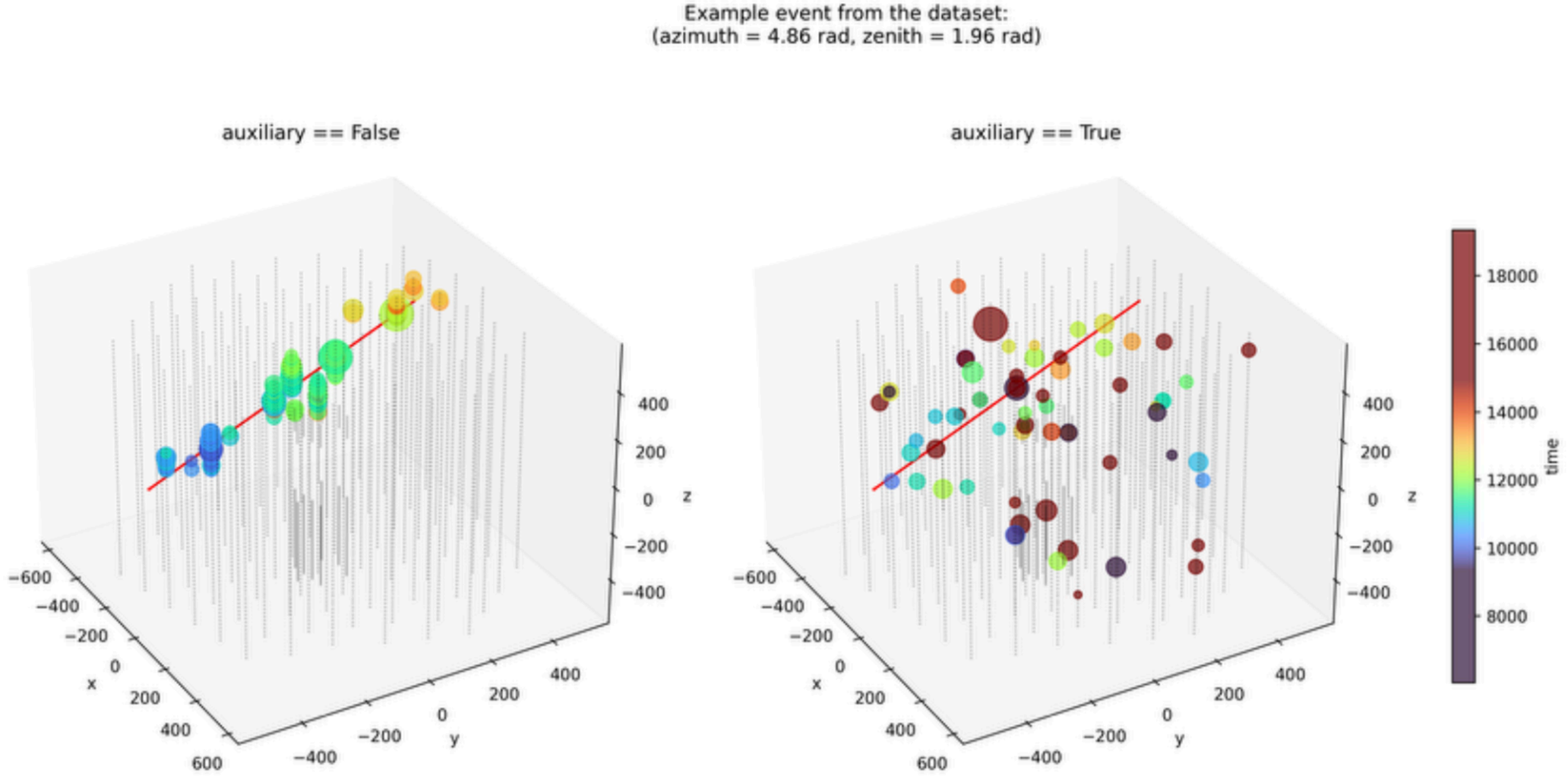
Usually lossy discretization:
- VQ-VAE (2401.13537, 2403.05618)
- pixelization (2402.10239)
- diffusion (2409.12589)

Challenges of self-supervise learning in particle physics

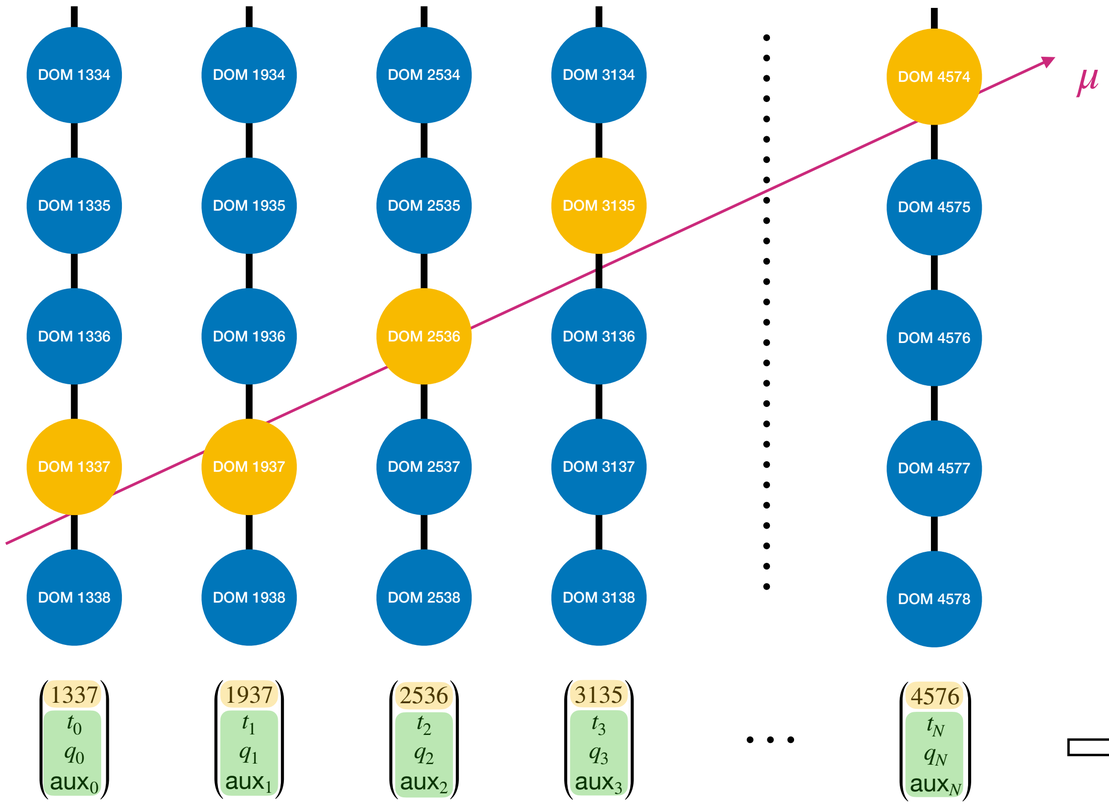
- How to predict a continuous 4-vector?
- Usually lossy discretization:
  - VQ-VAE (2401.13537, 2403.05618)
  - pixelization (2402.10239)
- How to sort 4-vectors?
- IceCube
  - 5160 DOMs — natural “tokenization”
  - Pulses have timestamps



IceCube event

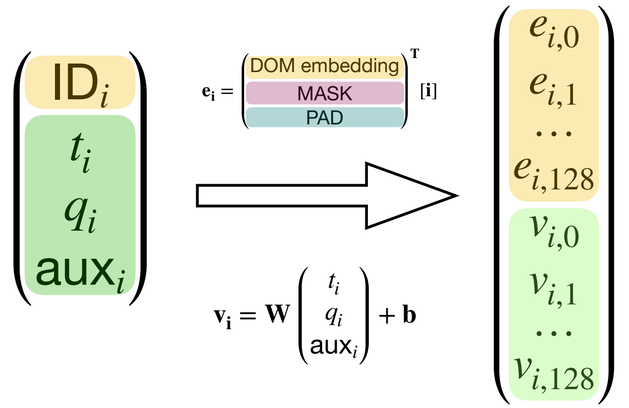


IceCube Embedding

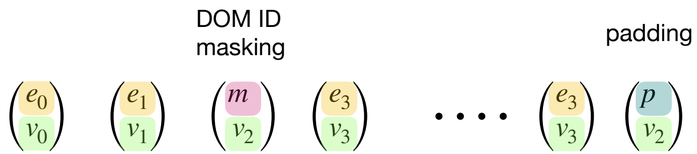


pulses (arranged by time)

pulse embedding

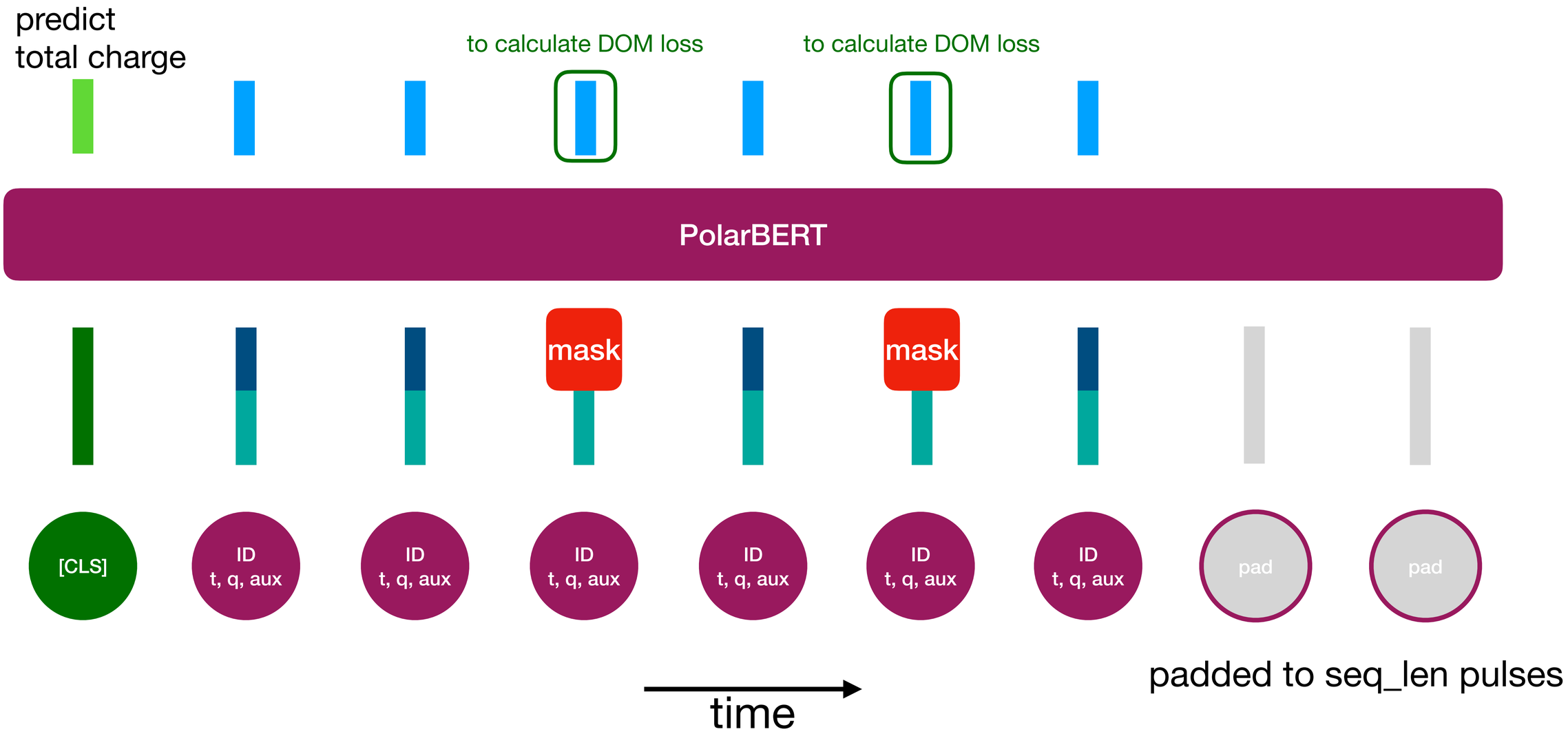


No position data!



time-series (padded to fixed length)

Pretraining

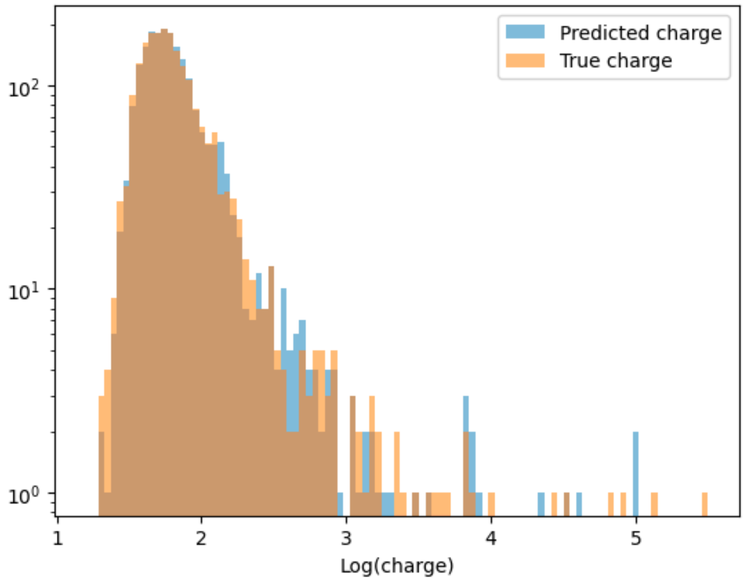


Pretraining: DOM loss

- The detection process is inherently stochastic
- We cannot predict the next DOM with certainty
- Similarly to LLMs, we use cross-entropy
  (but other option are possible: Earth Mover's Distance, Chamfer distance)
- DOM-loss: $L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_i)$, the sum over N masked doms
- Use only aux=false (HLC) pulses! aux=true pulses are impossible to predict.

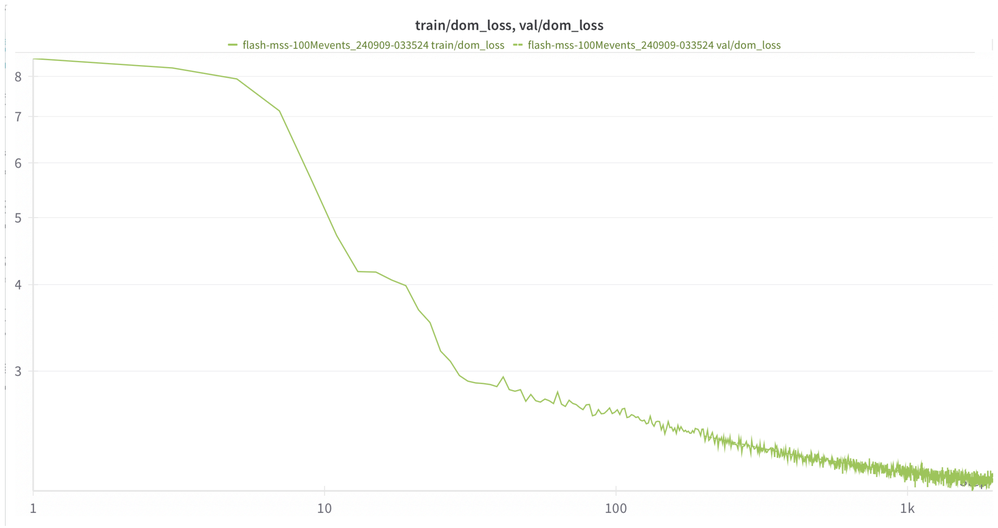
Pretraining: regression loss

- The model has to learn how to collect useful information in [CLS] embedding for the future use on downstream tasks.
- We need some feature that is not directly accessible to the model, but can be obtained from the data (no labels)
- Candidates: the total charge of the event, center of charge
- We subsample the events, and the charge is provided as a log
- Charge prediction loss: MSE($\log$(total charge))



PolarBERT: Foundation Model For IceCube

- Backbone: transformer (could be GRU, Mamba)
- Pretraining:
  - Subsample events to seq_len (currently 128)
  - input: (DOM embedding) $\oplus$ (projection of features)
  - loss function = DOM-loss + $\lambda \times$ charge-prediction-loss
- Fine-tuning for downstream tasks
- IceCube kaggle MC data for both pretraining and finetuning (studies using real data can be only published by the collaboration)

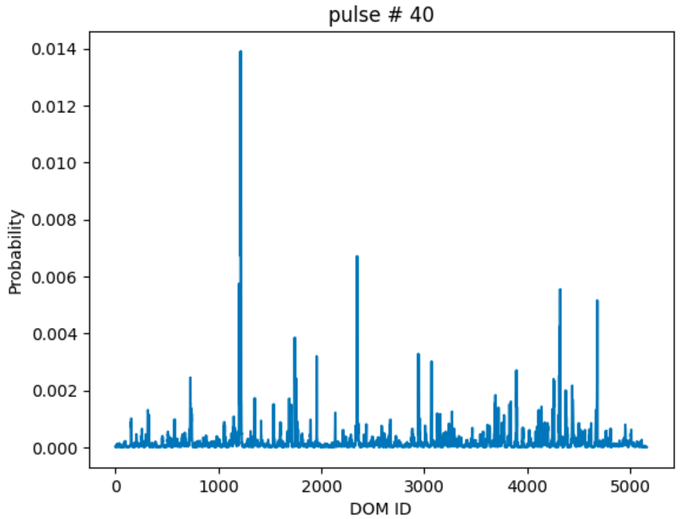


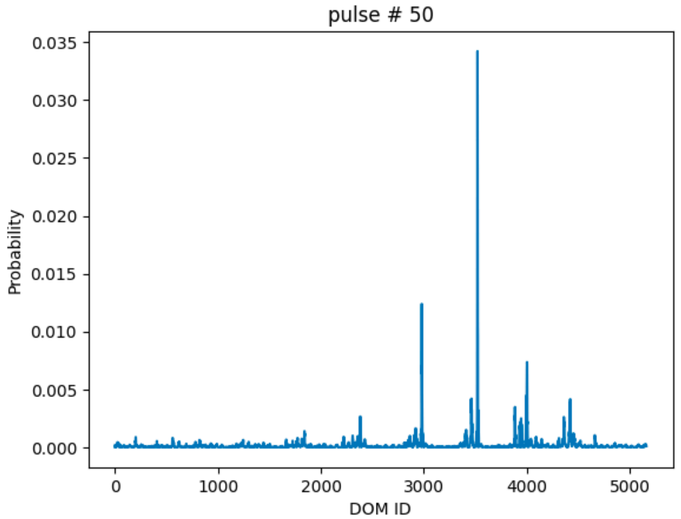
100 million events

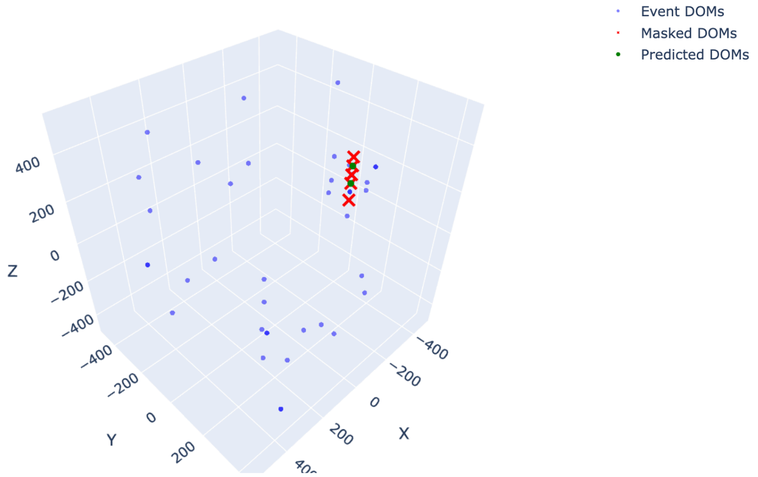
BERT: 3,300M tokens
PolarBERT: 127,000M “tokens”
(100M events x 127 pulses)

Interpreting the DOM Loss

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\log(p_i)$$



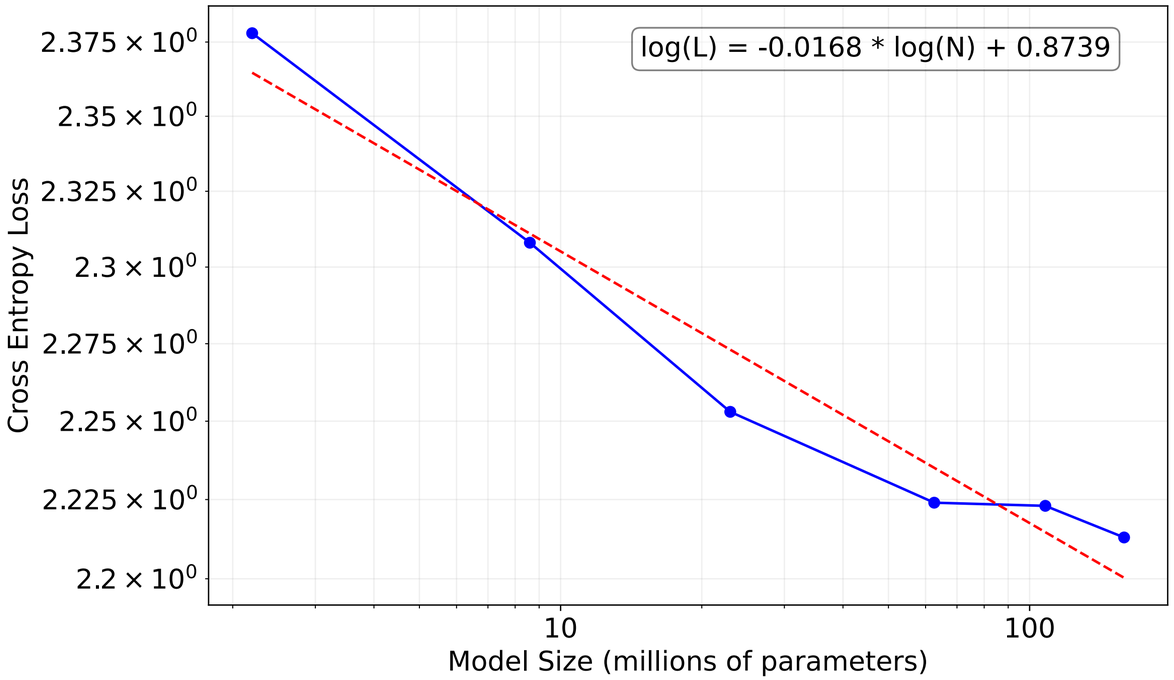




some uncertainty about the string and the DOM

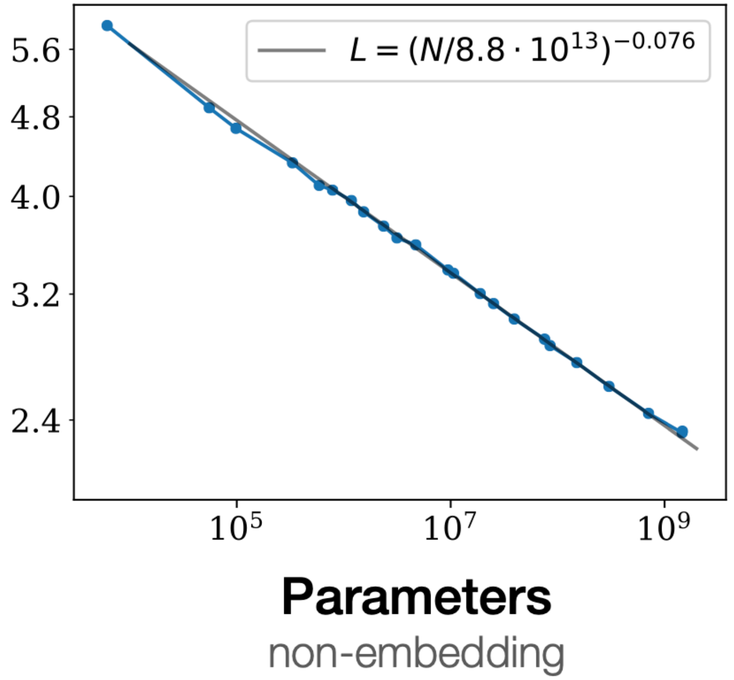
Model Size Scaling

PolarBERT



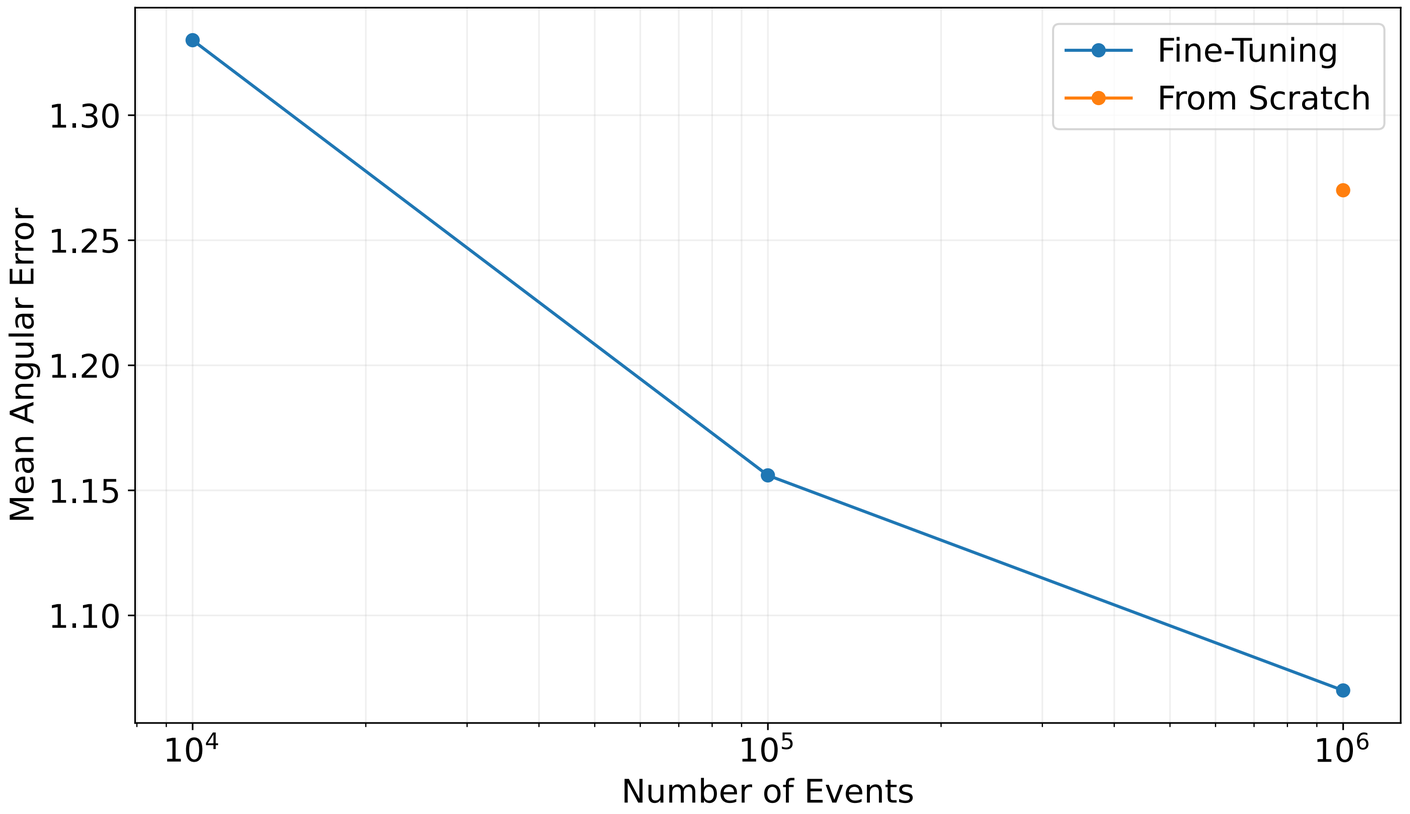
Models trained on 10M neutrino events

LLMs



Models trained to convergence
Kaplan et all, 2020

Finetuning (Directional Reconstruction)



- Pretrained model can be successfully fine-tuned on a downstream task
- We add a “prediction head”: an MLP to the [CLS] embedding output
- Train resulting model with direction labels
- Fine-tuning is sample efficient
- Allows to experiment with the architecture of the fine-tuned model

But what is a Transformer?

Attention is all you need

A Vaswani, N Shazeer, N Parmar… - Advanc

The dominant sequence transduction models
orconvolutional neural networks in an encod
performing such models also connect the en
echanisms. We propose a novel, simple netv
mechanism, dispensing with recurrence and
machine translation tasks show these model

☆ Save 99 Cite Cited by 74101 Relate

May 2023

Attention is all you need

A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, AN Gomez, Ł Kaiser, I Polosukhin

Advances in neural information processing systems, 2017 • proceedings.neurips.cc

Abstract

The dominant sequence transduction models are based on complex recurrent orconvolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attentionm echanisms. We propose a novel, simple network architecture based solely onan attention mechanism, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superiorin quality while being more

SHOW MORE

☆ Save 99 Cite Cited by 186030 Related articles All 73 versions »

June 2025

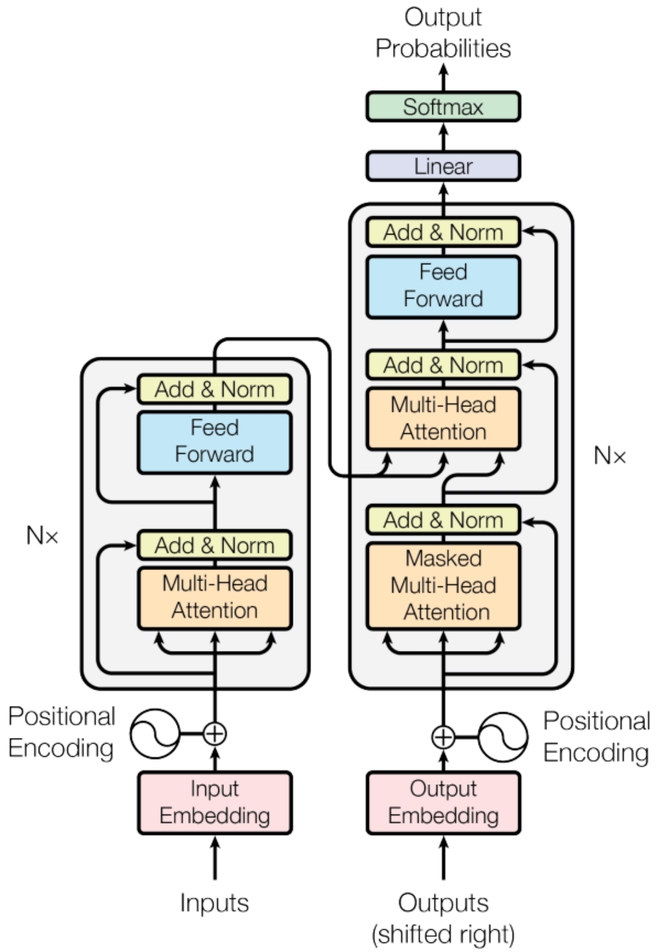


Figure 1: The Transformer - model architecture.

But what is a Transformer?

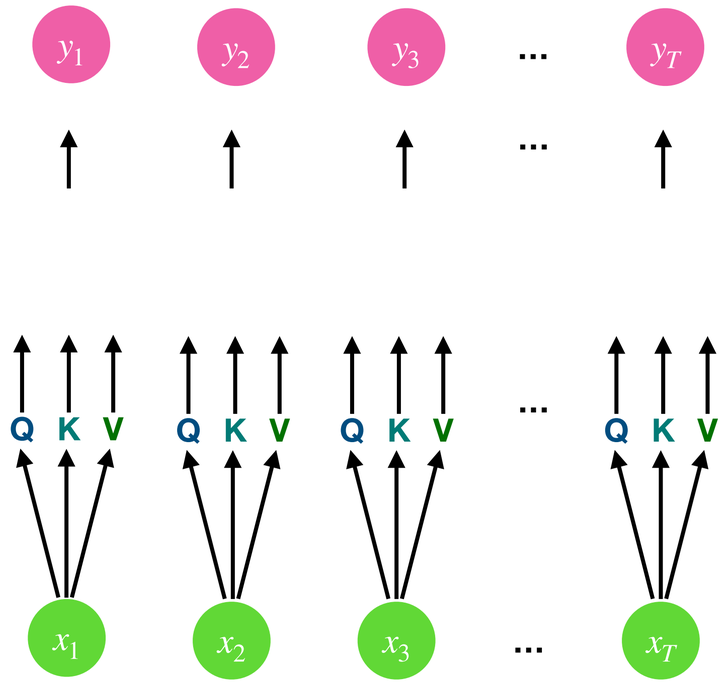
- A map from a set of inputs $\{ x_i \}$ to a set of outputs $\{ y_i \}$
- Attention mechanism is permutationally equivariant (hence sets)
- x_i could be vector representations of words
- y_i could be probabilities of words

- One can train a simple transformer model in 45 minutes on Google Colab (Here is an example on Shakespeare texts)

Attention Mechanism

- $x_{t\,i}$ a vector representation of a word. i - coordinates of the vector, t - its position in the text. (typical dimensionality 1000-10,000)
- Attention outputs: $y_{t\,i} = A_{ts}(x)\, W_{ij}\, x_{s\,j}$
- Attention matrix: $A_{ts}(x) = \frac{1}{Z} \exp(x_{t\,i}\, G_{ij}\, x_{s\,j})$
- $A_{ts}(x)$ depends on inputs. It also grows with the size of the text
- Both W_{ij} and G_{ij} are low rank (rank typically 64 or 128)

Attention Mechanism



$$y_t = \sum_s \text{softmax}\left(\frac{(W^Q x_t) \cdot (W^K x_s)^T}{\sqrt{d_K}}\right)(W^V x_s)$$

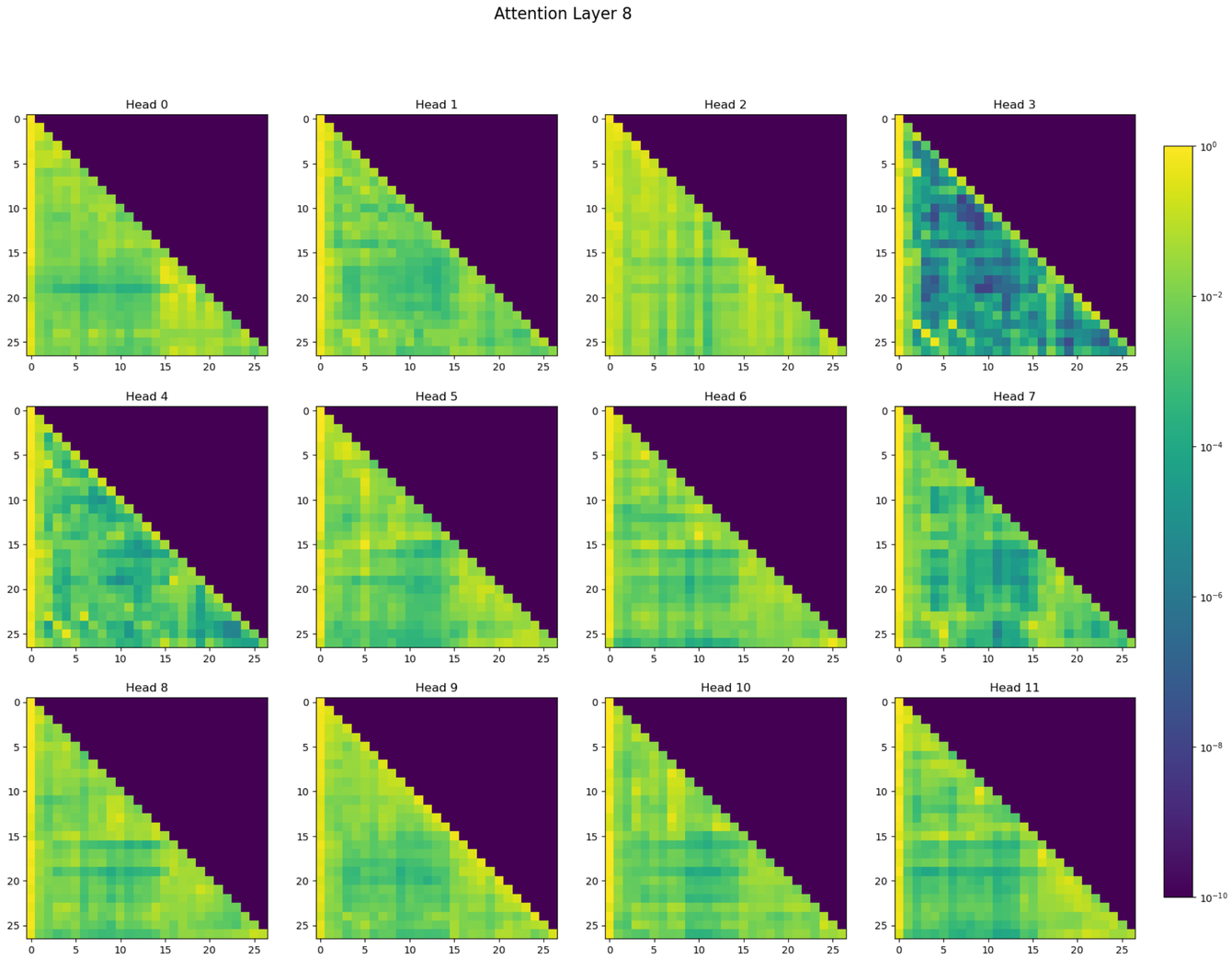
- Attention outputs:

$$y_{t\,i} = A_{ts}(x)\, W_{ij}\, x_{s\,j}$$

- Attention matrix:

$$A_{ts}(x) = \frac{1}{Z} \exp(x_{t\,i}\, G_{ij}\, x_{s\,j})$$

How Self-Attention works?



… Nobody knows

but some progress
see e.g. https://transformer-circuits.pub/

Transformer Model

- A full transformer model includes:
- A matrix mapping words form a dictionary to vectors (each row corresponds to a word). Typical size ~100,000 x 10,000
- Normalization — projecting the vector onto a sphere
- Repeated blocks of:
  - Attention (multiple in parallel): $y_{t\,i} = A_{ts}(x)\, W_{ij}\, x_{s\,j}$, $A_{ts}(x) = \frac{1}{Z} \exp(x_{t\,i}\, G_{ij}\, x_{s\,j})$

    (also something to break permutation equivariance)
  - Multi-Layer Perceptron: $y_i = W_{ia} \max(W_{aj} x_j, 0)$
- A map of the outputs to probabilities over the dictionary $p_d = \frac{1}{Z} \exp\left(\frac{l_d}{T}\right)$

LLMs, Foundation Models and Physics

- Broadly, there are two directions:
  - Specialized Models (e.g. AlphaFold, Nobel Prize in chemistry 2024)
  - Generic Models (ChatGPT, Gemini, Claude)

- Can we use benefit from the generic models?
- Are foundation models useful?
  (Adding visual capabilities to LLMs does’t improve their text capabilities)

LLMs, Foundation Models and Physics

- LLMs become better and better:

  Quantum Many-Body Physics Calculations with Large Language Models
  https://arxiv.org/abs/2403.03154
  “We evaluate GPT-4's performance in executing the calculation for 15 research papers from the past decade, demonstrating that, with correction of intermediate steps, it can correctly derive the final Hartree-Fock Hamiltonian in 13 cases and makes minor errors in 2 cases.”

- Recently*: “*AlphaEvolve: A coding agent for scientific and algorithmic discovery” https://arxiv.org/abs/2506.13131

LLMs, Foundation Models and Physics

Student

- Has some domain knowledge
- No knowledge of a specific project
- Starts from a blank state
- Can work on any project
- Has to defend a thesis
- Would I give a student a project with 1% success chance?

Gemini

- Has a lot of domain knowledge
- No knowledge of a specific project
- Starts from a blank state
- Can work on any project
- Doesn’t have to defend a thesis
- Would you give an LLM a project with 1% success chance?

Outlook

- We can leverage unlabeled data for IceCube direction reconstruction
- Scaling also works in physics (but with smaller exponents)

Backup Slides

Representing words

- Vocabulary: enumerate all words
  But there are too many words in many languages

- Tokenization: Breaking text into smaller units (words, sub-words)
  Note! Tokenization causes its own issues
  [very detailed lecture]

Tokens 220 Characters 747

```
Week 4 (Convolutional Neural Networks (CNNs), Recurrent Neural Networks (
RNNs), and Auto-Encoders (AE)):
May 15: 13:15-17:00: Convolutional Neural Networks (CNNs) and image
 analysis (Daniel Murnane).
    Exercise: Recognize images (MNIST dataset, sparse chips for
 radiation, and/or insoluables from Greenland ice cores) with a CNN.
May 17: 9:15-12:00: Recurrent Neural Networks (RNN), Long Short Term
 Memory (LSTM) and Natural Language Processing (NLP) (Inar Timiryasov).
    Exercise: Use an LSTM to predict flight traffic and do Natural
 Language Processing on IMDB movie reviews.
May 17: 13:15-17:00: (Variational) Auto-Encoder and anomaly detection (TP
).
    Exercise: Compress images using Auto-Encoder, and cluster latent
 space with UMAP.
```

TEXT TOKEN IDS

A helpful rule of thumb is that one token generally corresponds to ~4 characters of text for common English text. This translates to roughly ¾ of a word (so 100 tokens ~= 75 words).

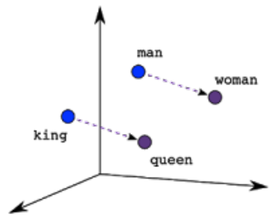
Tokens 220 Characters 747

```
[20916, 604, 357, 3103, 85, 2122, 282, 47986, 27862, 357, 18474, 82, 828,
3311, 6657, 47986, 27862, 357, 49, 6144, 82, 828, 290, 11160, 12, 4834,
19815, 364, 357, 14242, 8, 2599, 220, 198, 6747, 1315, 25, 1511, 25,
1314, 12, 1558, 25, 405, 25, 34872, 2122, 282, 47986, 27862, 357, 18474,
82, 8, 290, 2939, 3781, 357, 19962, 337, 700, 1531, 737, 198, 220, 220,
220, 220, 32900, 25, 31517, 1096, 4263, 357, 39764, 8808, 27039, 11,
29877, 12014, 329, 11881, 11, 290, 14, 273, 35831, 84, 2977, 422, 30155,
4771, 21758, 8, 351, 257, 8100, 13, 198, 6747, 1596, 25, 860, 25, 1314,
12, 1065, 25, 405, 25, 3311, 6657, 47986, 27862, 357, 49, 6144, 828,
5882, 10073, 35118, 14059, 357, 43, 2257, 44, 8, 290, 12068, 15417,
28403, 357, 45, 19930, 8, 357, 818, 283, 5045, 9045, 292, 709, 737, 198,
220, 220, 220, 220, 32900, 25, 5765, 281, 406, 2257, 44, 284, 4331, 5474,
4979, 290, 466, 12068, 15417, 28403, 319, 8959, 11012, 3807, 8088, 13,
198, 6747, 1596, 25, 1511, 25, 1314, 12, 1558, 25, 405, 25, 357, 23907,
864, 8, 11160, 12, 27195, 12342, 290, 32172, 13326, 357, 7250, 737, 198,
220, 220, 220, 220, 32900, 25, 3082, 601, 4263, 1262, 11160, 12, 27195,
12342, 11, 290, 13946, 41270, 2272, 351, 471, 33767, 13]
```

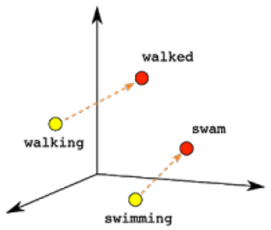
TEXT TOKEN IDS

https://platform.openai.com/tokenizer

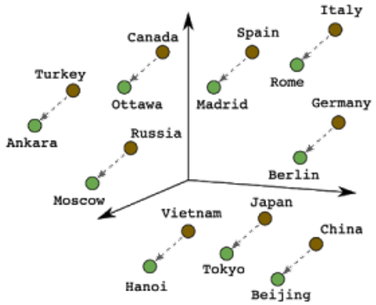
- Embeddings — every token is a vector in a multidimensional space
  (Word2Vec)



Male-Female



Verb Tense

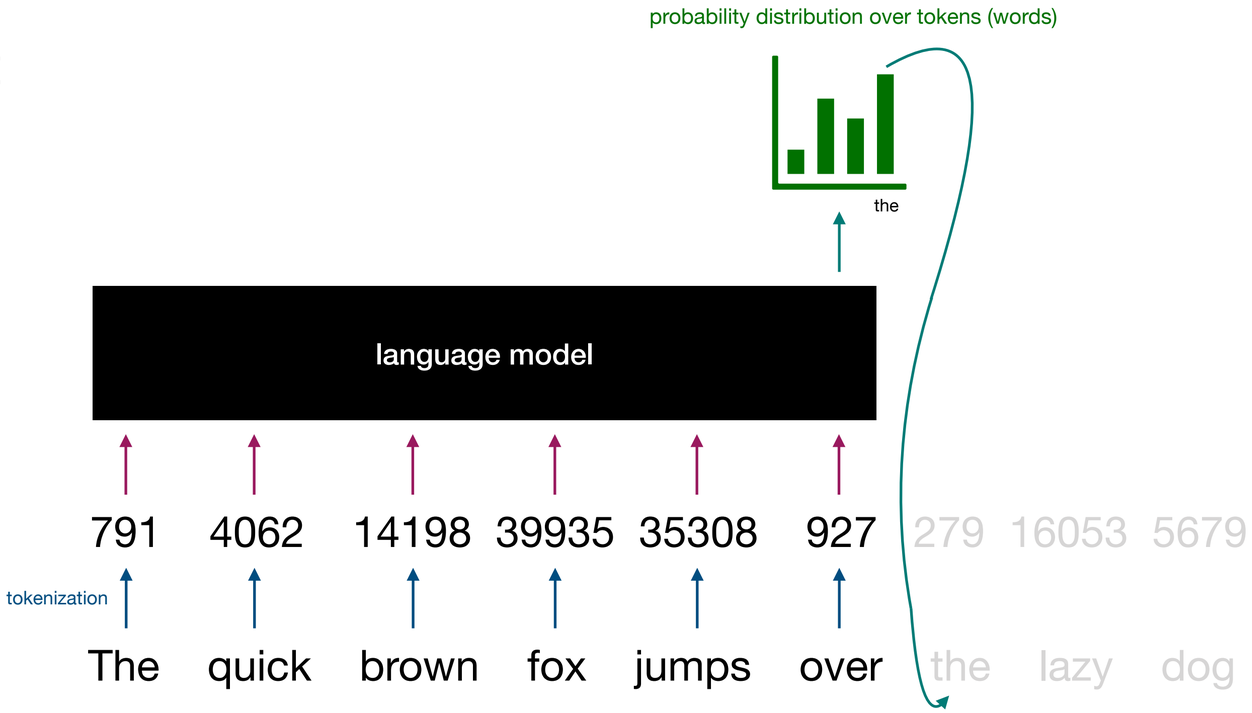


Country-Capital

Operations over vectors: `king - man + woman ~= queen`

Language modeling: generic picture

- Process the first t words
- Predict the probability of the next word $P(x^{(t+1)} \mid x^{(1)}, x^{(2)}, \ldots, x^{(t)})$
- Sample the next word
- Process the first $t + 1$ words
- …

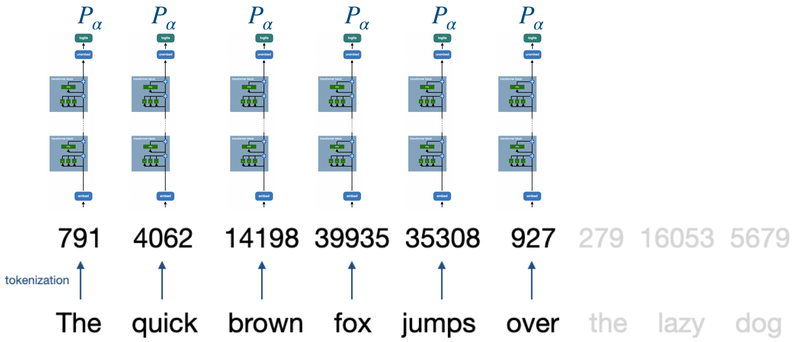


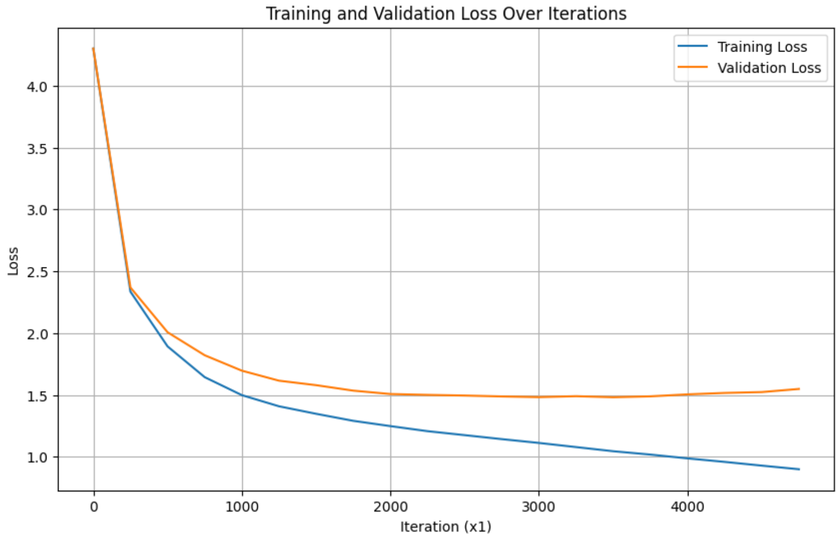
How to train an LLM

- **Self-supervised pretraining**
- Output probabilities over all tokens (words)

$$P_\alpha = softmax(\ell_\alpha) = \frac{\exp(\ell_\alpha)}{\sum_\beta \exp(\ell_\beta)}, \text{ where } \ell_\alpha - \text{logits (raw outputs)}$$

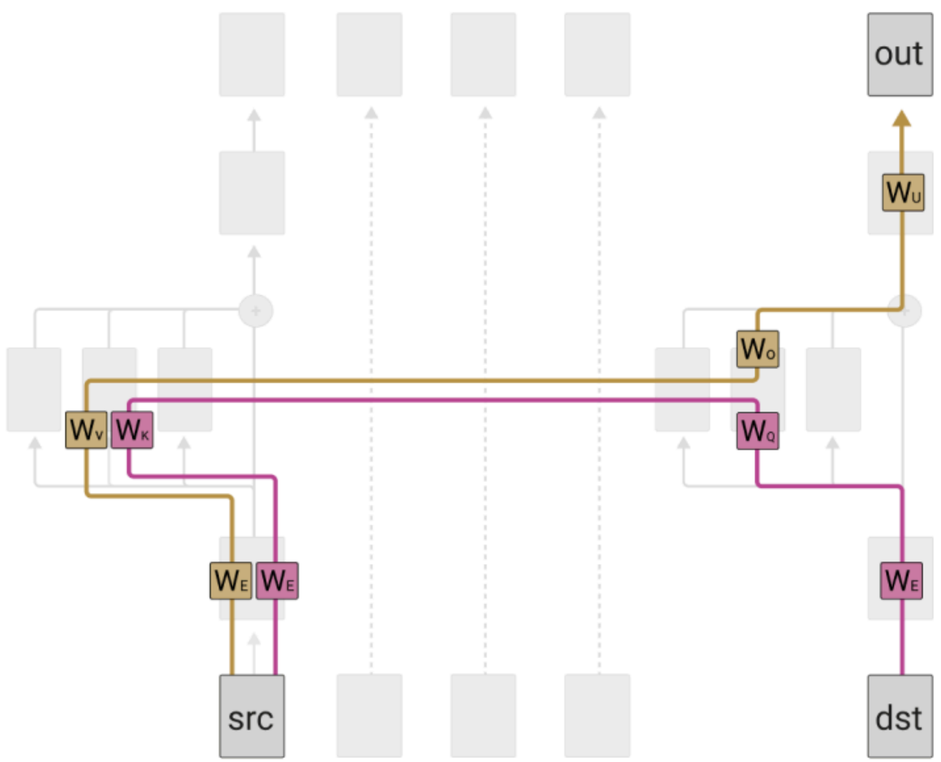
- Cross-Entropy Loss $= -\sum_s \sum_\alpha y_\alpha \log(P_\alpha)$, where y_α are true next tokens (one-hot encoded).
- Typically trained with very large batch sizes (millions of tokens). Highly parallelizable.
- Adam optimizer (or similar). SGD doesn't work for transformers!





nanoGPT exercise!

How attention works?



The **OV ("output-value") circuit** determines how attending to a given token affects the logits.

$$W_U W_O W_V W_E$$

The **QK ("query-key") circuit** controls which tokens the head prefers to attend to.

$$W_E^T W_Q^T W_K W_E$$

Source https://transformer-circuits.pub/2021/framework/index.html