Understanding Foundation Model
Theory & Algorithms

Sungbin Lim
Department of Statistics, College of Political Science and Economics

2026. 04. 24 전남대

Department of Statistics
KOREA UNIVERSITY

KOREA
UNIVERSITY

Ice Breaking

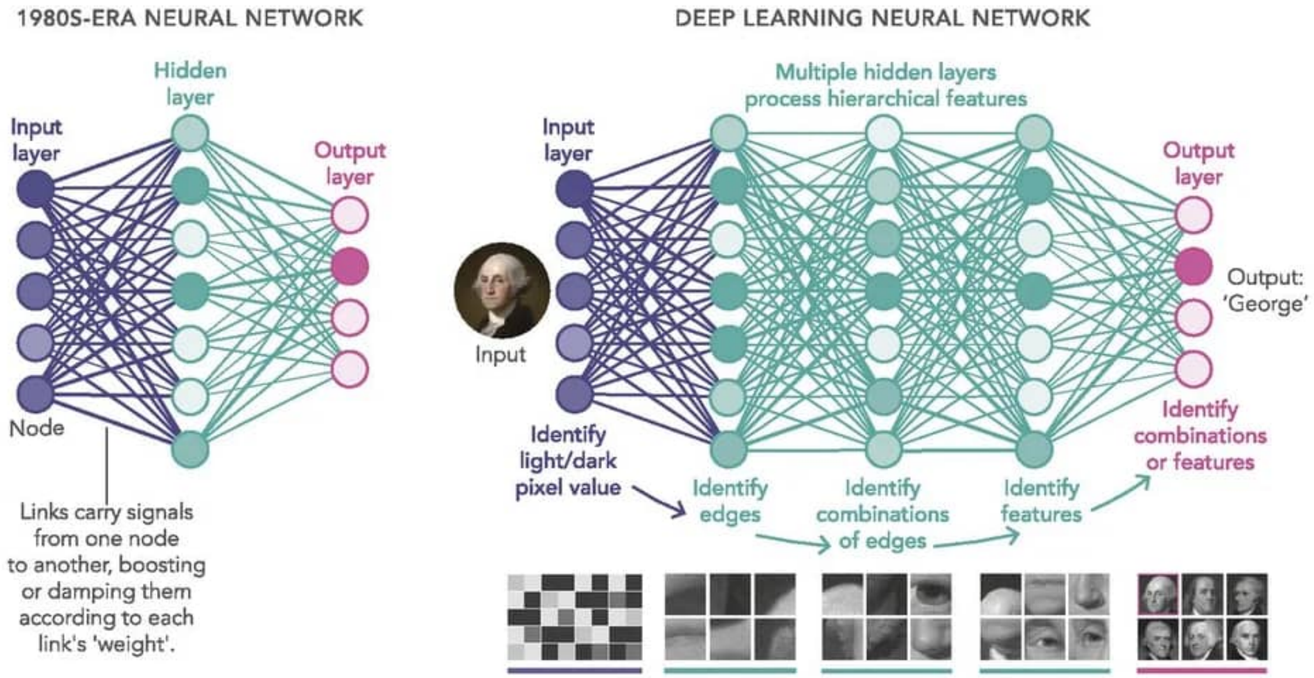
- **Lim, Sungbin** (林聖彬)

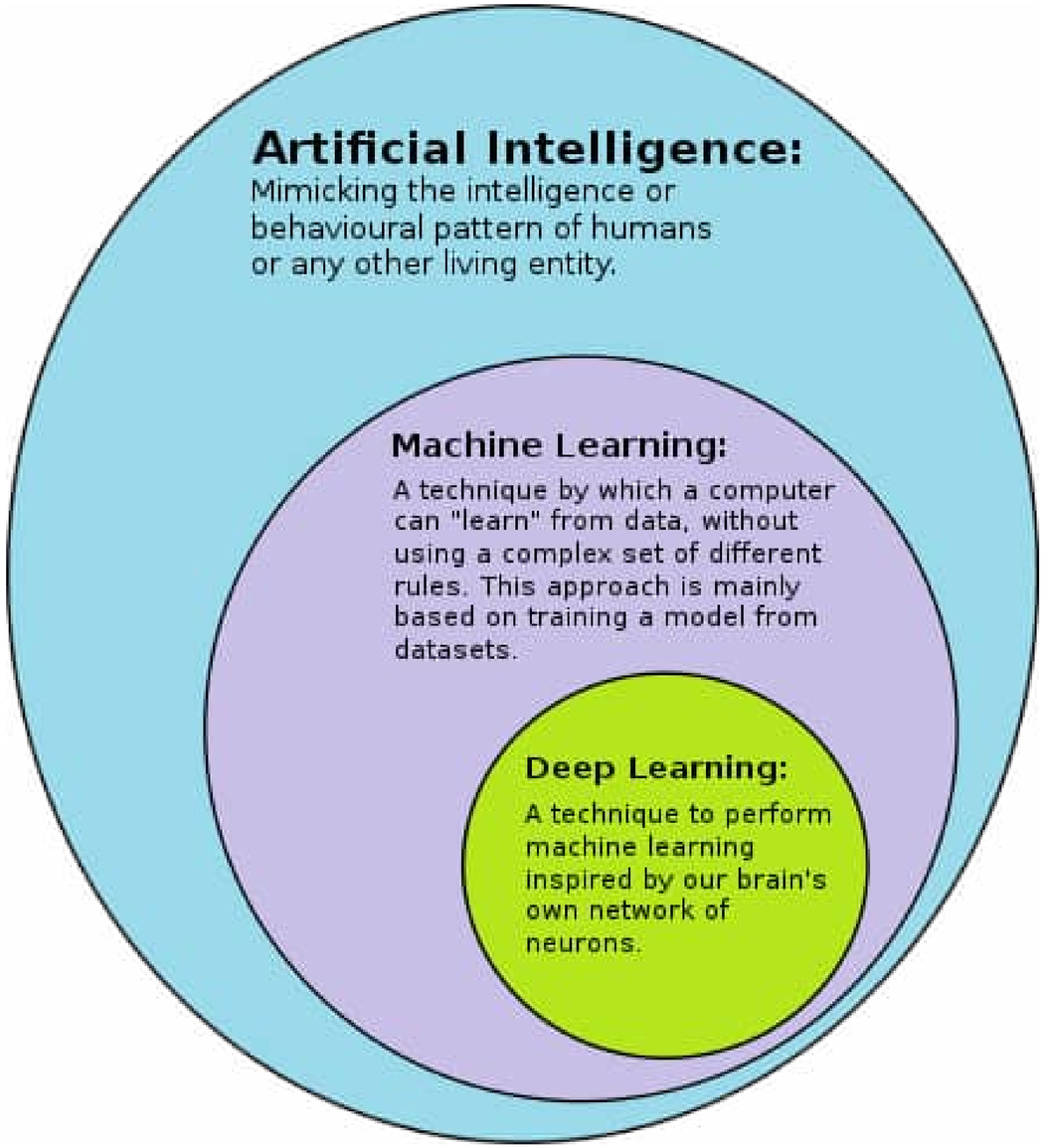
  - Ph.D. in Mathematics
  - Associate Professor of Department of Statistics at KOREA UNIVERSITY
  - Collaborate Research w/ LG AI Research

- **Research Area**
  - Artificial Intelligence
  - Stochastic Optimization
  - Probabilistic Machine Learning & Reasoning

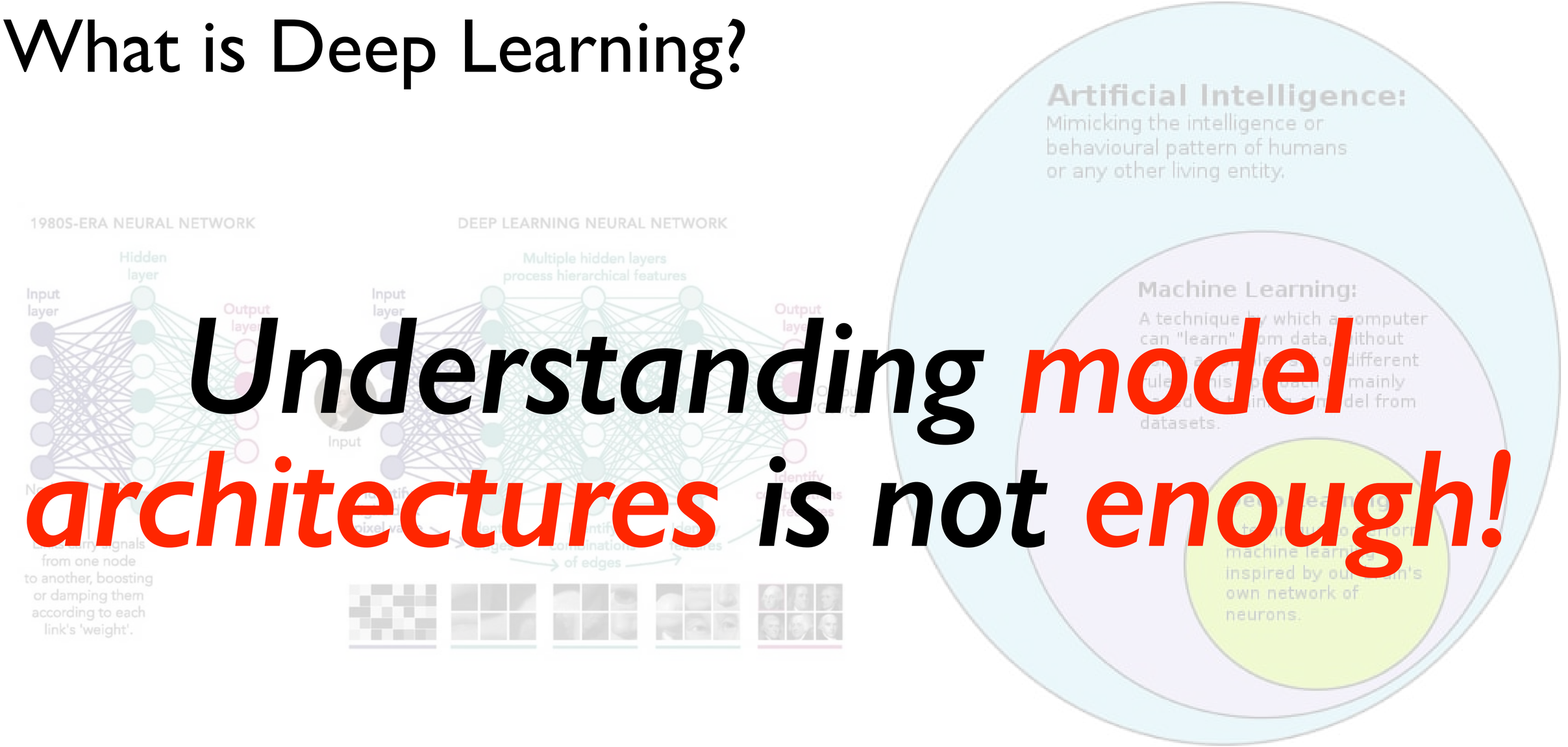
Part I
Toward Machine Reasoning

What is Deep Learning?



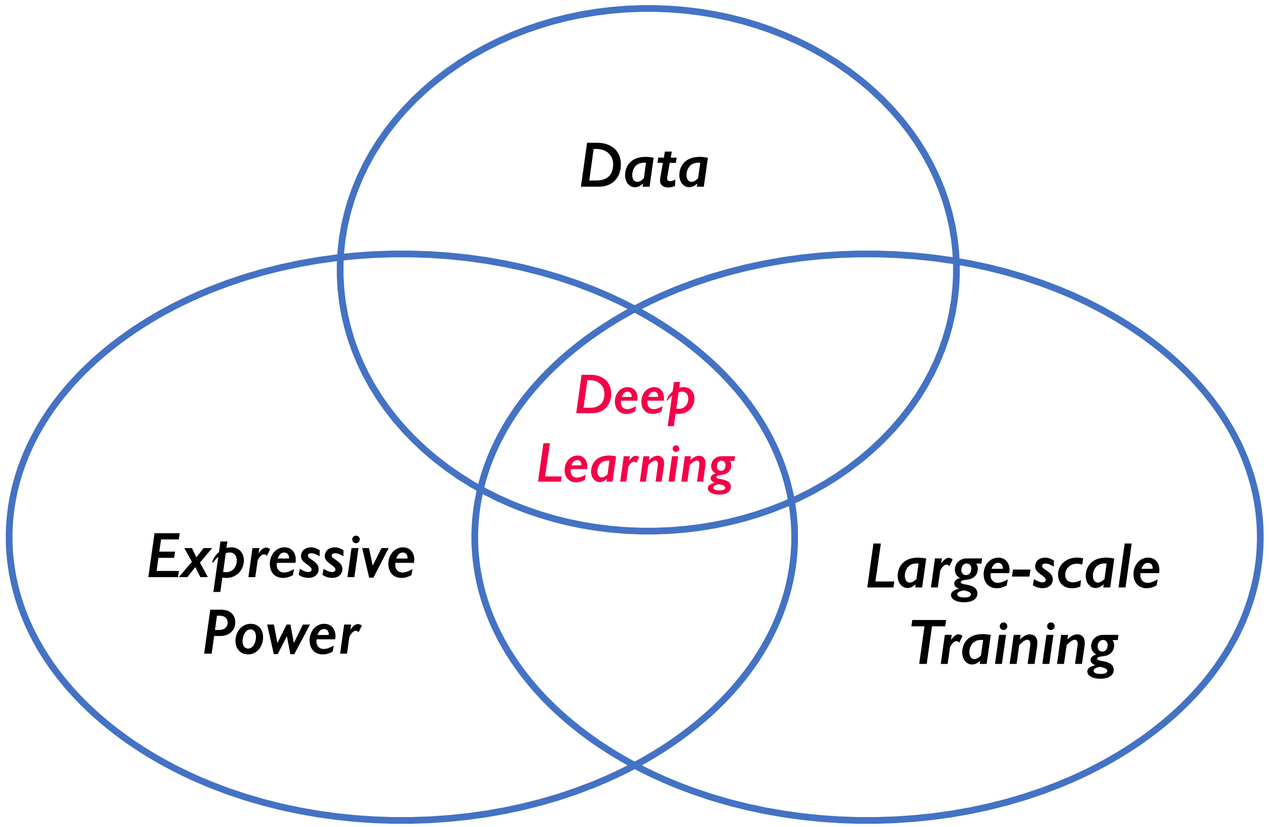


What is Deep Learning?

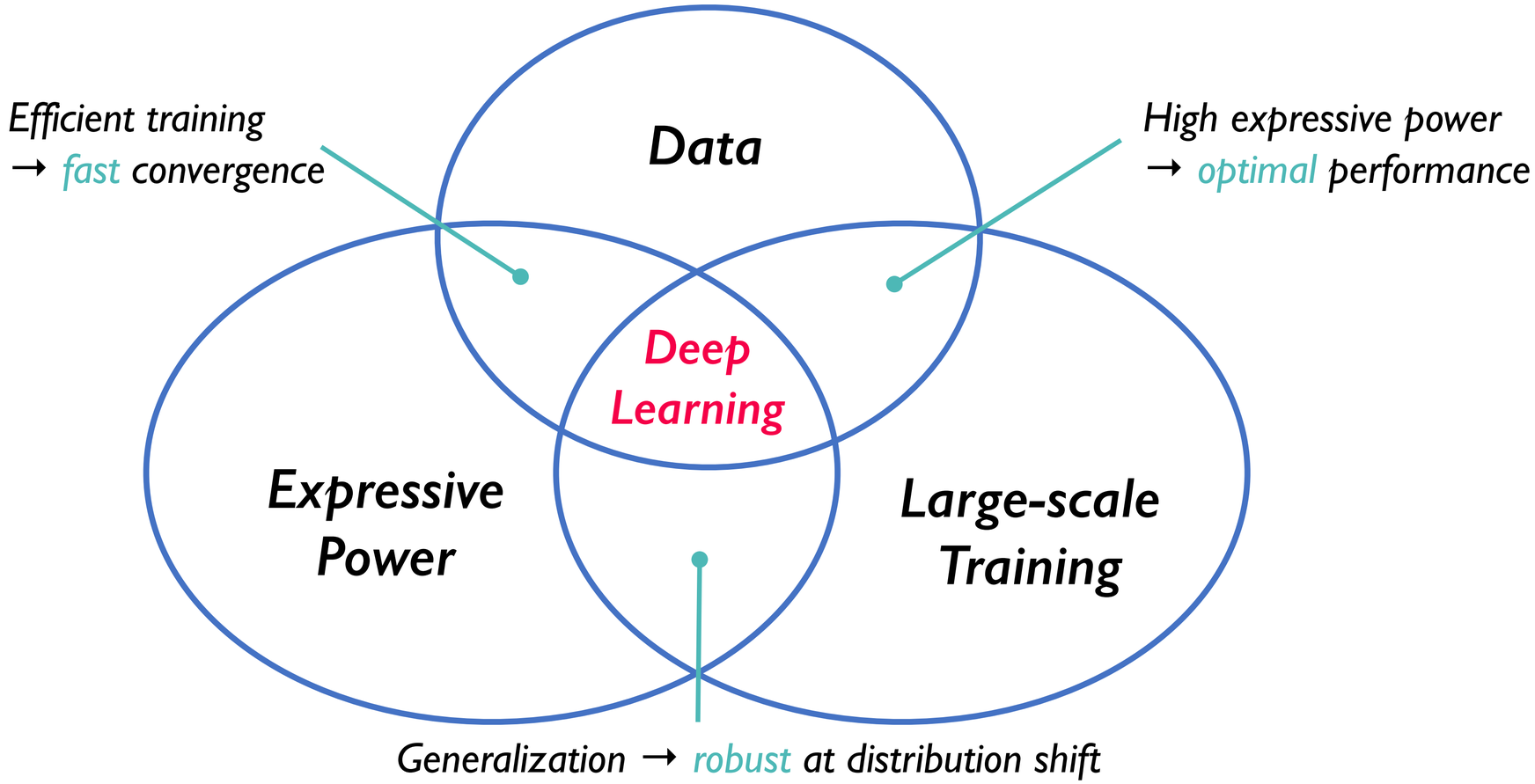


Understanding model architectures is not enough!

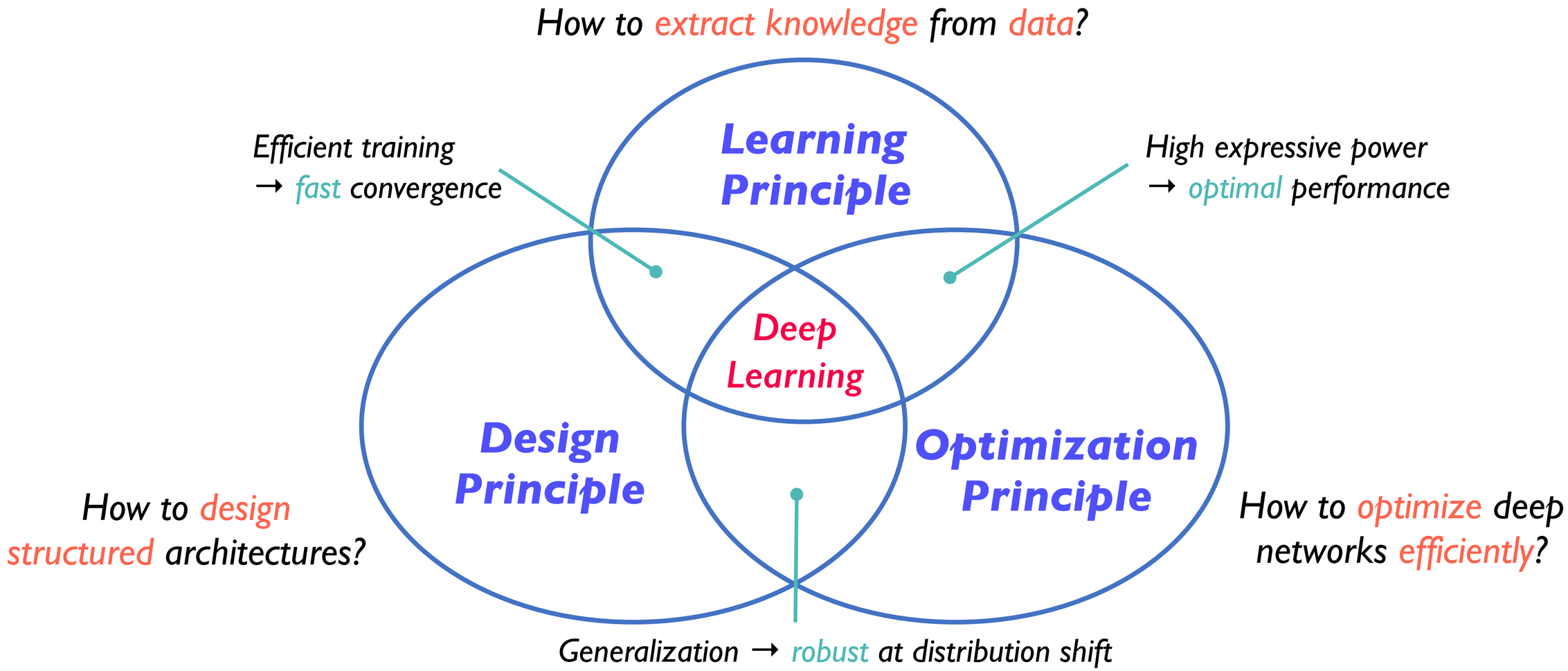
Three Principles of Deep Learning



Three Principles of Deep Learning

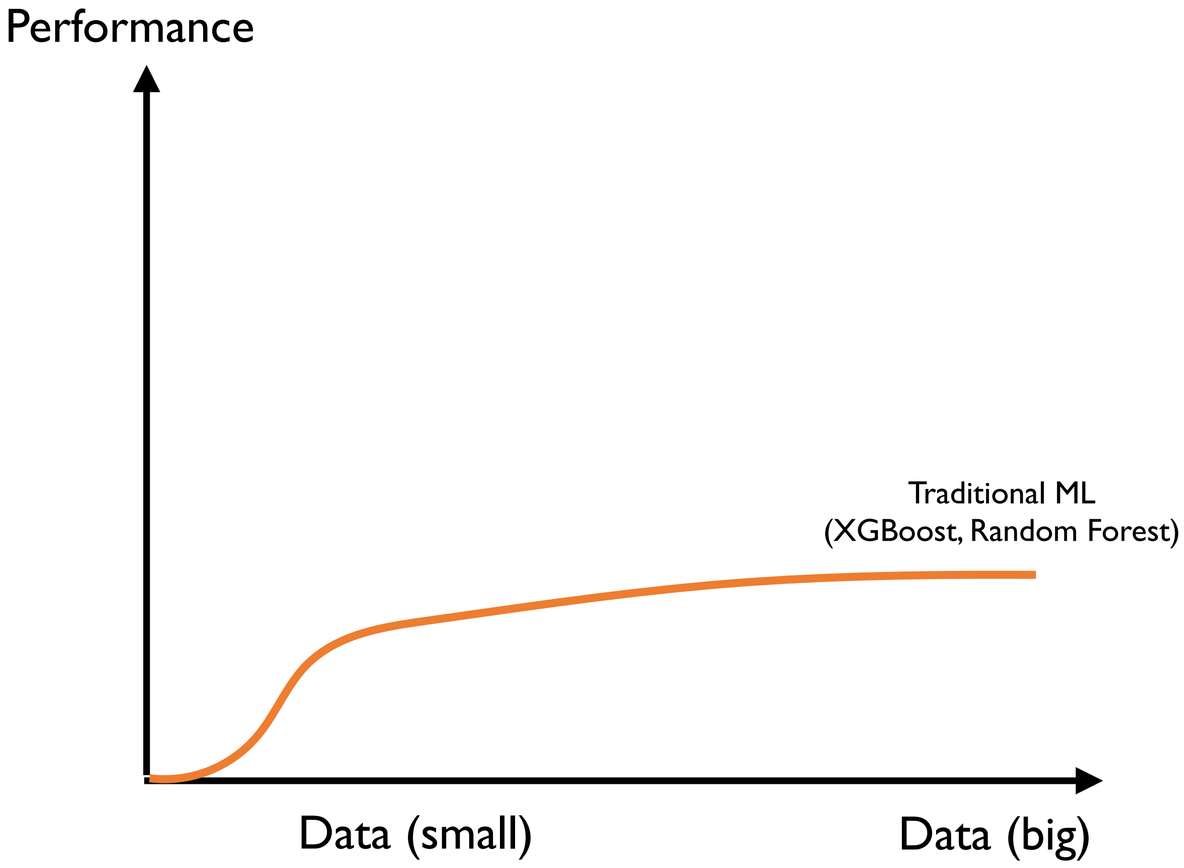


Three Principles of Deep Learning



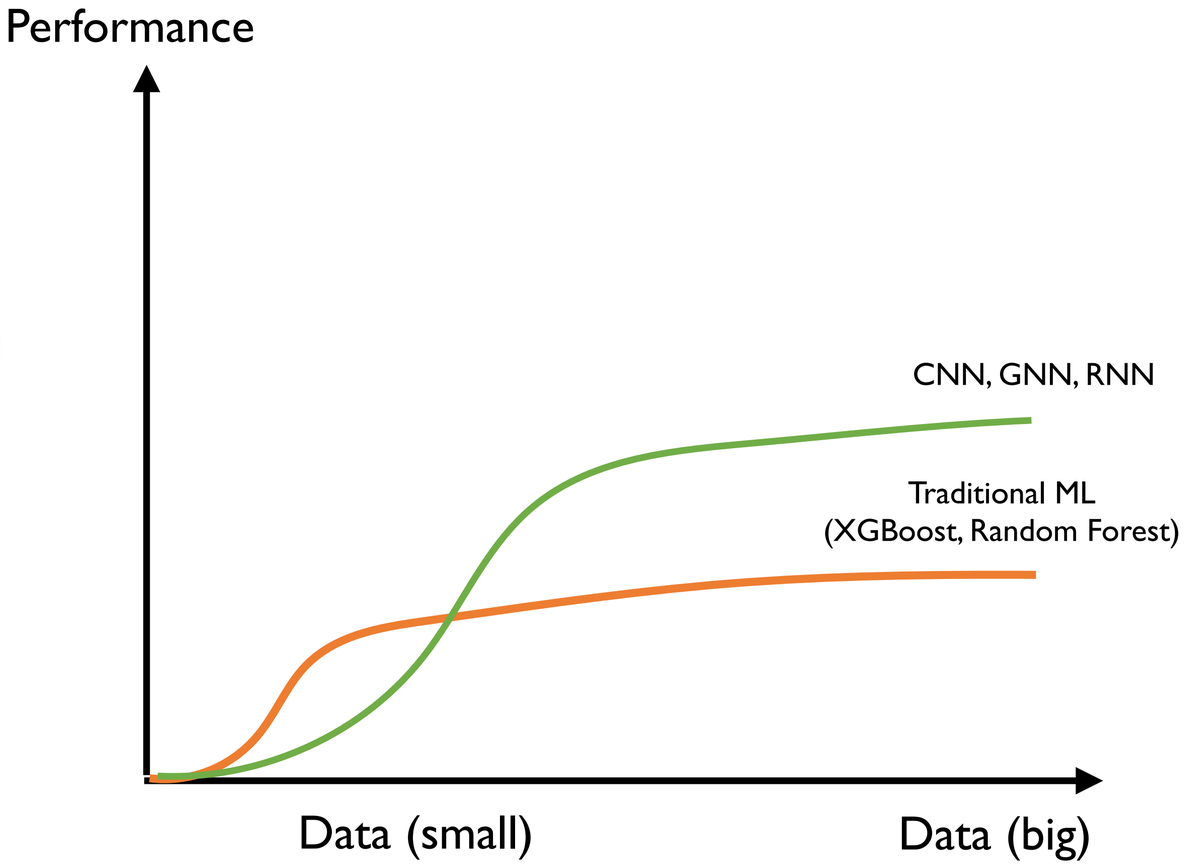
Deep Learning = 💸

- If you have small number of data, then traditional ML performs well



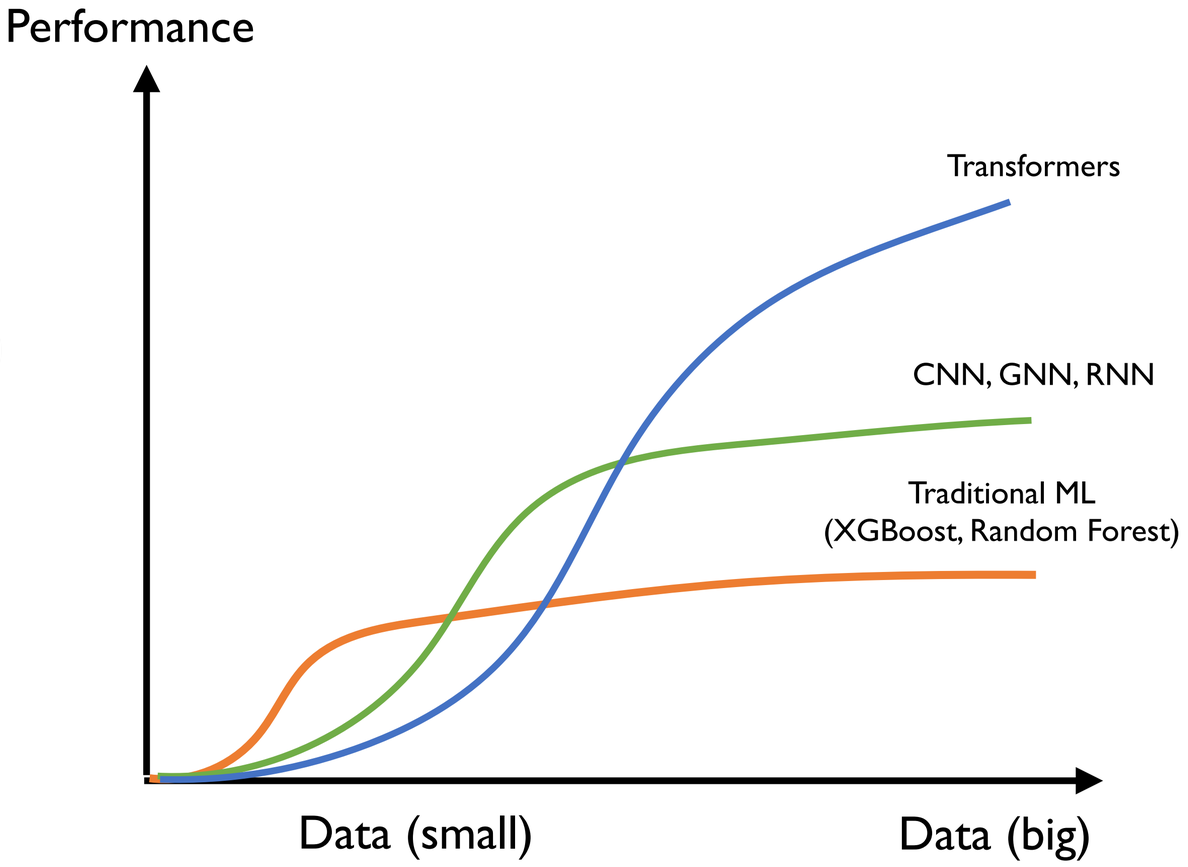
Deep Learning = 💸

- If you have small number of data, then traditional ML performs well
- If you have more data, then you can try deep learning with well-designed inductive biases (e.g. CNN, GNN, RNN)



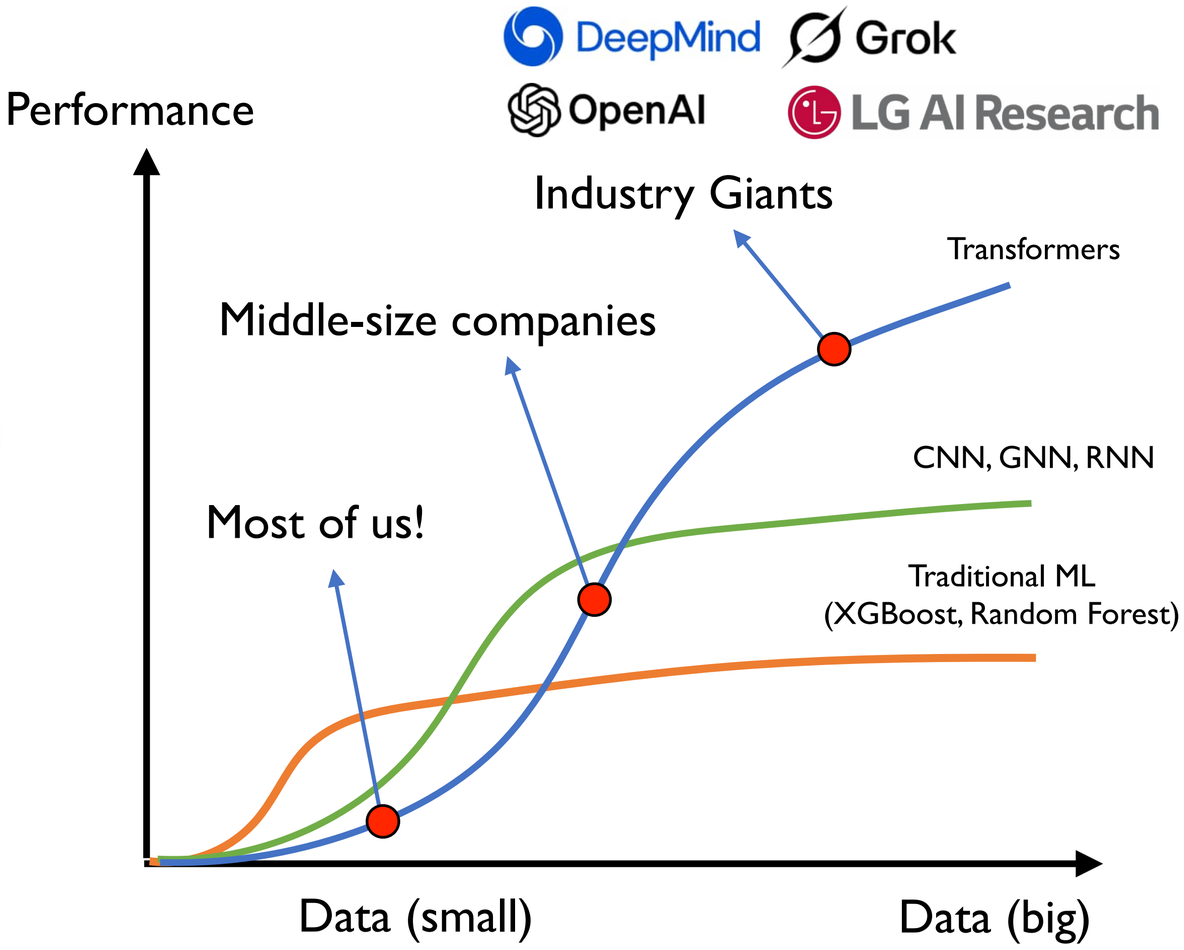
Deep Learning = 💸

- If you have small number of data, then traditional ML performs well
- If you have more data, then you can try deep learning with well-designed inductive biases (e.g. CNN, GNN, RNN)
- If you have large number of data and sufficient computing resources, then you need to try Transformer-type (LLMs) architectures with a pretraining strategy



Deep Learning = 💸

- If you have small number of data, then traditional ML performs well
- If you have more data, then you can try deep learning with well-designed inductive biases (e.g. CNN, GNN, RNN)
- If you have large number of data and sufficient computing resources, then you need to try Transformer-type (LLMs) architectures with a pretraining strategy
- Learning from scratch is ***meaningless***



Artificial Intelligence

Artificial Intelligence

Machine Learning

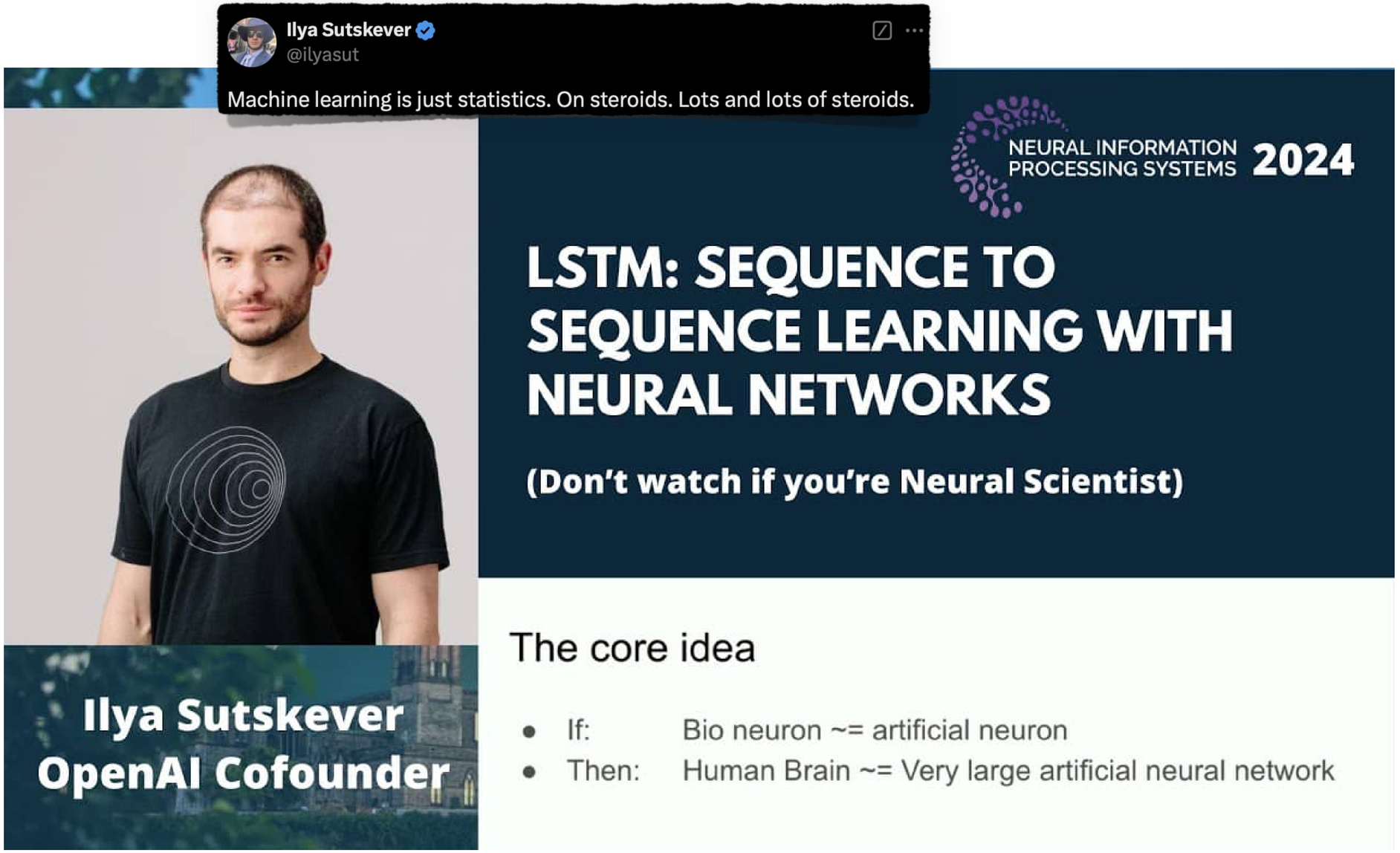
everybody knows! 😪

Artificial Intelligence

Machine Learning

Machine Reasoning

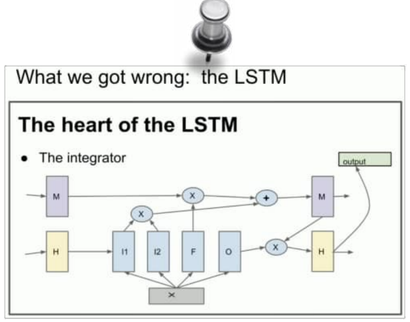
What is this? 🤔



Source: NeurIPS 2024 Test-time Talk

Learning Principle of Future?

What we got right: Deep Learning / Autoregressive Models (Transformer) / Scaling Hypothesis



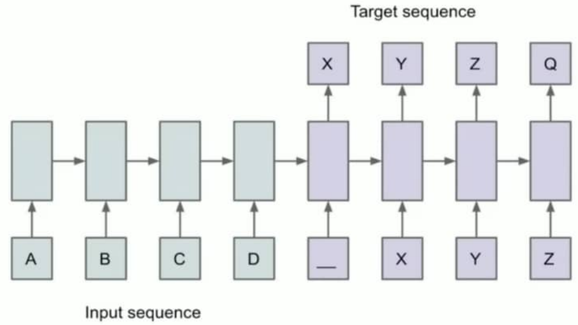
"The Deep Learning Hypothesis"

- Human perception is fast
  - Neurons fire at most 100 times a second
  - Humans solve perception in 0.1 seconds

  → our neurons fire 10 times, at most

Anything a human can do in 0.1 seconds, a big 10-layer neural network can do, too!

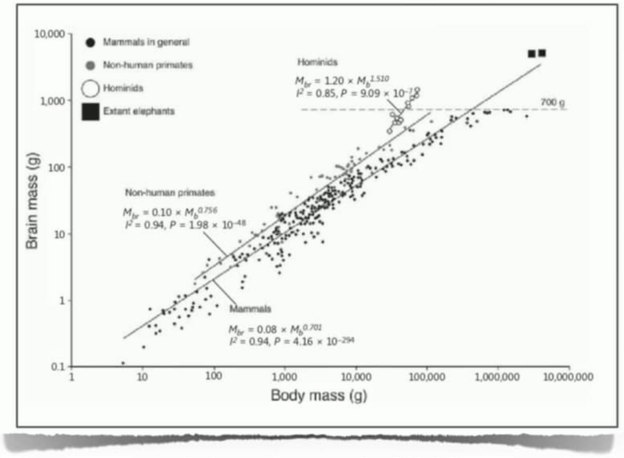
Our main idea



Conclusions

- If you have a large big dataset
- And you train a very big neural network
- Then success is guaranteed!

What comes next? Example from nature

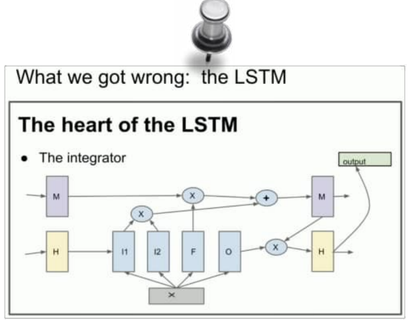


Manger et al., 2013

Ilya Sutskever, Seq2seq Learning with Neural Networks: what a decade, ***NeurIPS 2024 Test-time Talk***

Learning Principle of Future?

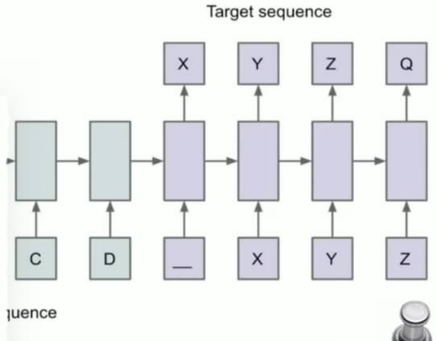
What we got right: Deep Learning / Autoregressive Models (Transformer) / Scaling Hypothesis



"The Deep Learning Hypothesis"

- Human perception is fast
  - Neurons fire at most 100 times a second
  - Humans solve perception in 0.1 seconds

Our main idea



Conclusions

- If you have a large big dataset
- And you train a very big neural network
- Then success is guaranteed!

Pre-training as we know it will end

Compute is growing:

- Better hardware
- Better algorithms
- Larger clusters

Data is not growing:

- We have but one internet
- **The fossil fuel of AI**

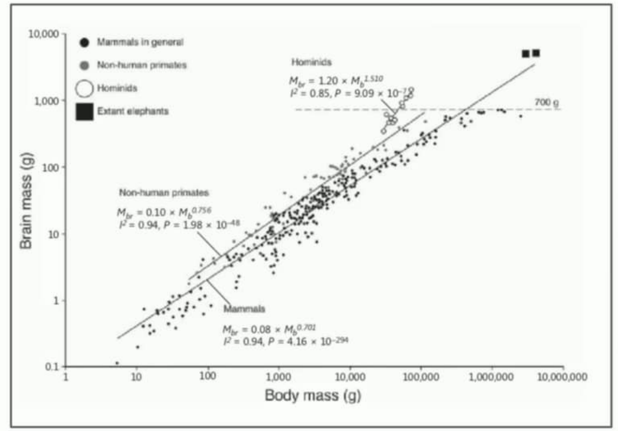
What comes next? The long term

Superintelligence

- Agentic
- Reasons
- Understands
- Is self aware

How humans do Science?

What comes next? Example from nature



Manger et al., 2013

Ilya Sutskever, Seq2seq Learning with Neural Networks: what a decade, *NeurIPS 2024 Test-time Talk*

Machine *Learning* vs Machine *Reasoning*

- **Machine Learning** focuses on mining the hidden patterns from data to tackle a pre-determined problem

supervised learning
(e.g. classification, regression) $\{(x_i, y_i)\} \rightarrow f_\theta$

unsupervised learning
(e.g. density estimation) $\{x_i\} \rightarrow g_\phi$

index of data samples

L. Bottou, From Machine Learning to Machine Reasoning, ***Machine Learning*** **(2014)**

Machine *Learning* vs Machine *Reasoning*

- **Machine Learning** focuses on mining the hidden patterns from data to tackle a pre-determined problem
- **Machine Reasoning** implements thinking process as a computational system by manipulating acquired knowledge and data to answer a new question

supervised learning
(e.g. classification, regression)
$\{(x_i, y_i)\} \rightarrow f_\theta$

unsupervised learning
(e.g. density estimation)
$\{x_i\} \rightarrow g_\phi$

index of data samples

reasoning w/ knowledge
(e.g. meta learning)
$\{(\mathcal{D}_\tau, f_\tau)\} \rightarrow F_\Theta$

learned models

reasoning w/o knowledge
(e.g. scientific discovery)
$\{\mathcal{D}_\tau, \square\} \rightarrow G_\Phi$

index of tasks

acquired knowledge

L. Bottou, From Machine Learning to Machine Reasoning, ***Machine Learning*** **(2014)**

Machine *Learning* vs Machine *Reasoning*

- **Machine Learning** focuses on mining the hidden patterns from data to tackle a pre-determined problem
- **Machine Reasoning** implements thinking process as a computational system by manipulating acquired knowledge and data to answer a new question

supervised learning (e.g. classification, regression) $\{(x_i, y_i)\} \rightarrow f_\theta$

reasoning w/ knowledge (e.g. meta learning) $\{(\mathcal{D}_\tau, f_\tau)\} \rightarrow F_\Theta$

unsupervised learning (e.g. density estimation) $\{x_i\} \rightarrow g_\phi$

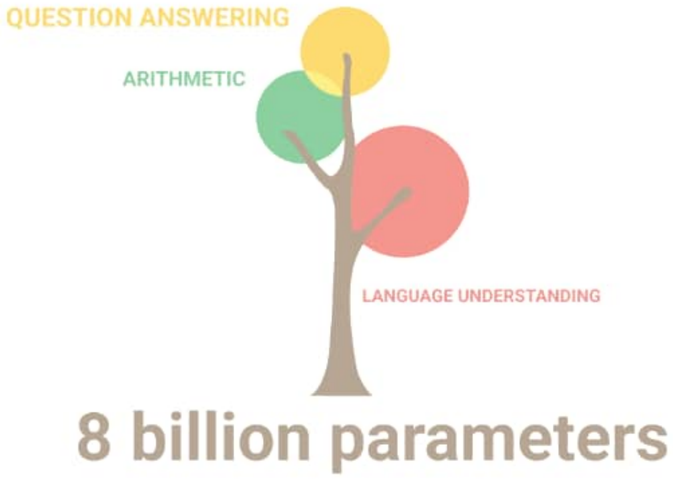
index of data samples

reasoning w/o knowledge (e.g. scientific discovery) $\{\mathcal{D}_\tau, \square\} \rightarrow G_\Phi$

How can we develop such computational system? 🤔

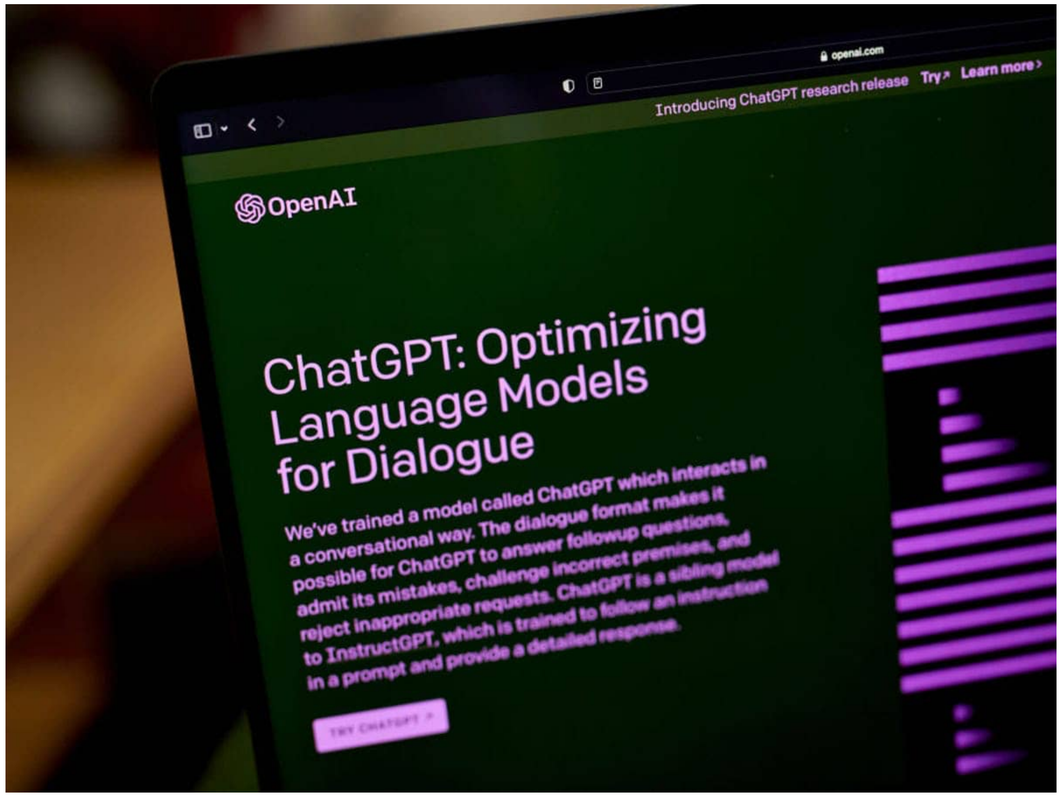
L. Bottou, From Machine Learning to Machine Reasoning, ***Machine Learning*** **(2014)**

Can LLMs Learn Multi-Tasks and Reasoning?



Google AI Blog (2022)

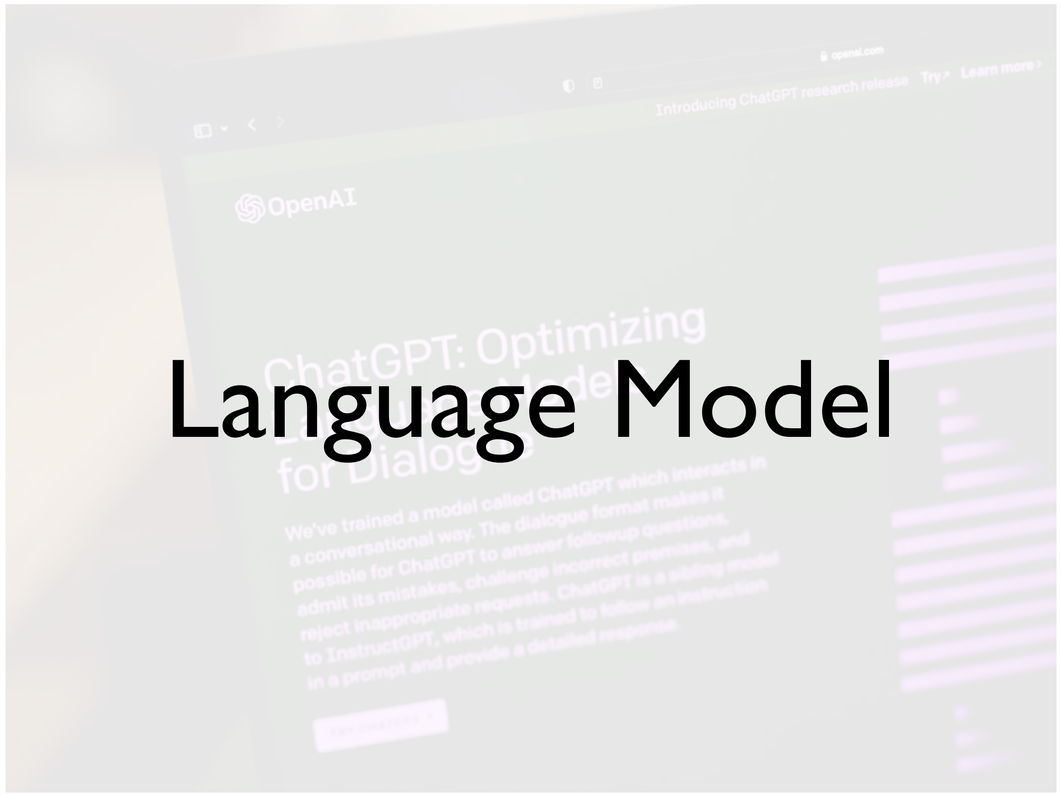
The Era of Generative AI



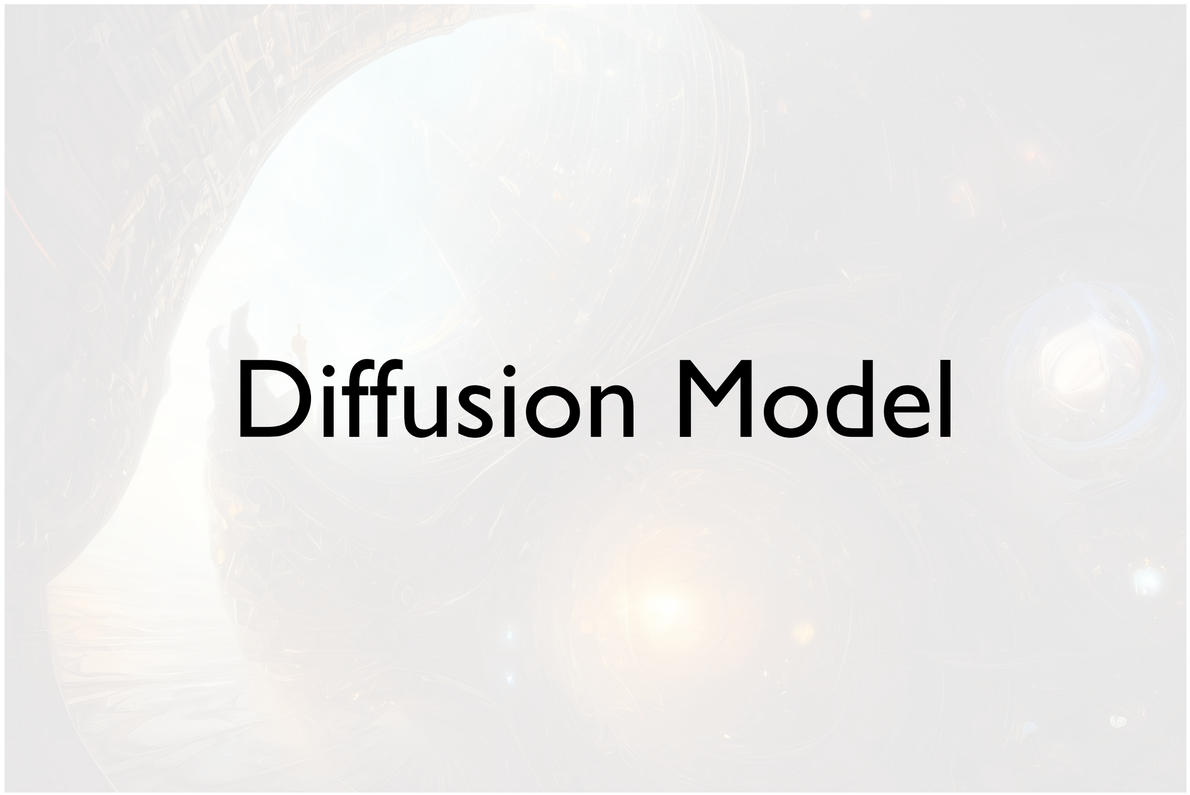
ChatGPT (Source: OpenAI)

Stable Diffusion (Source: Stability AI)

The Era of Generative AI



ChatGPT (Source: OpenAI)

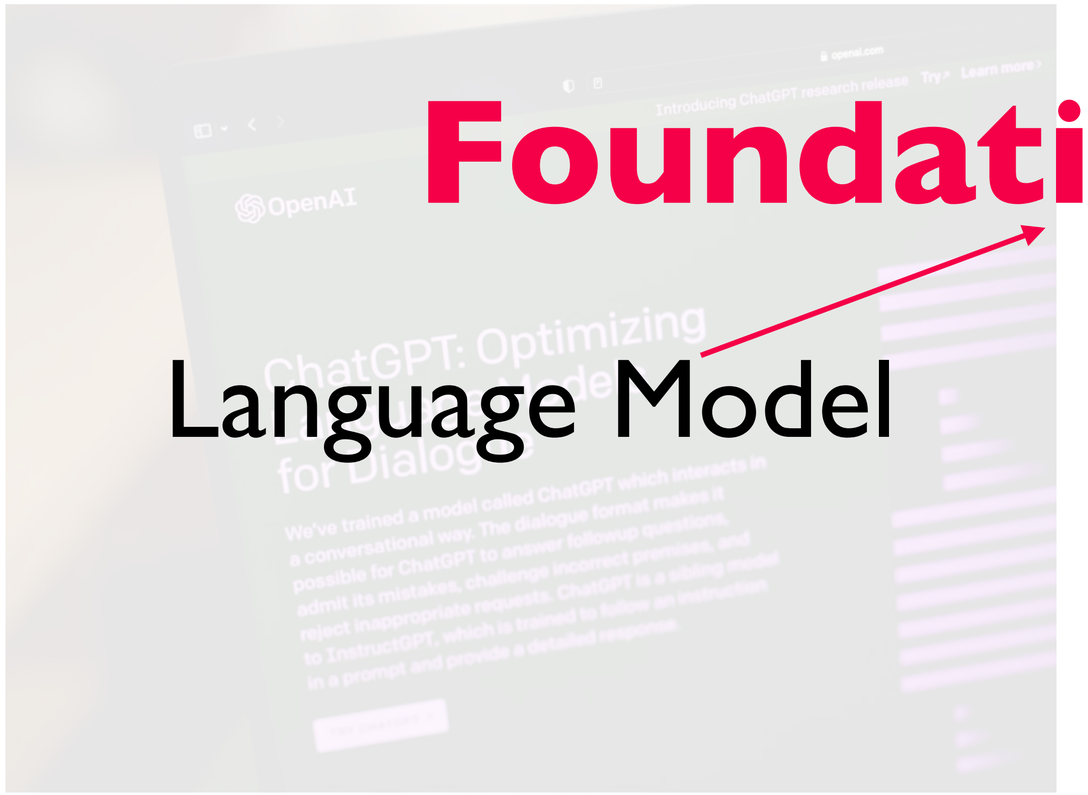


Stable Diffusion (Source: Stability AI)

The Era of Generative AI

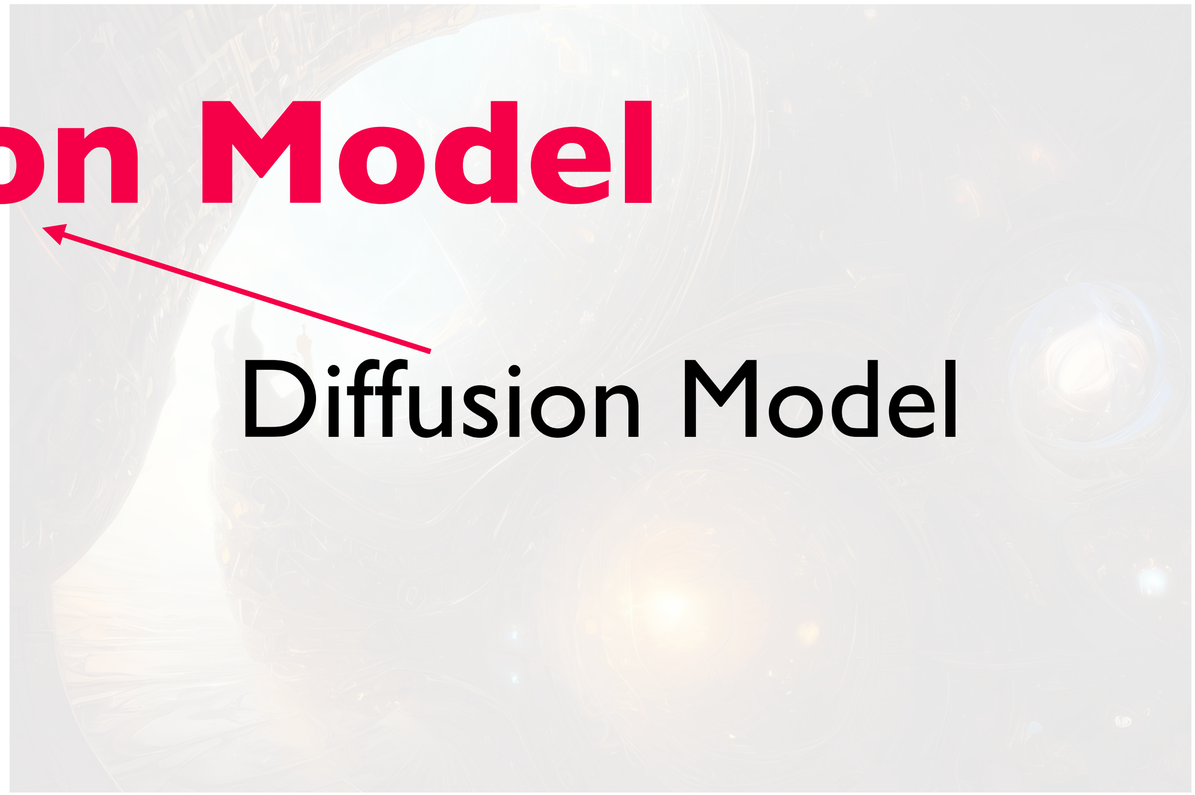
Foundation Model

Language Model



ChatGPT (Source: OpenAI)

Diffusion Model



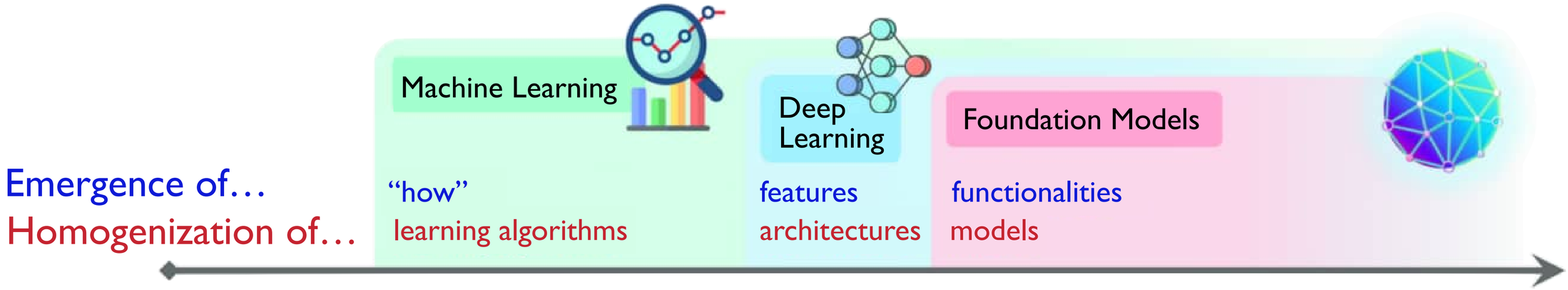
Stable Diffusion (Source: Stability AI)

Part 2

What is Foundation Model?

What is Foundation Model?

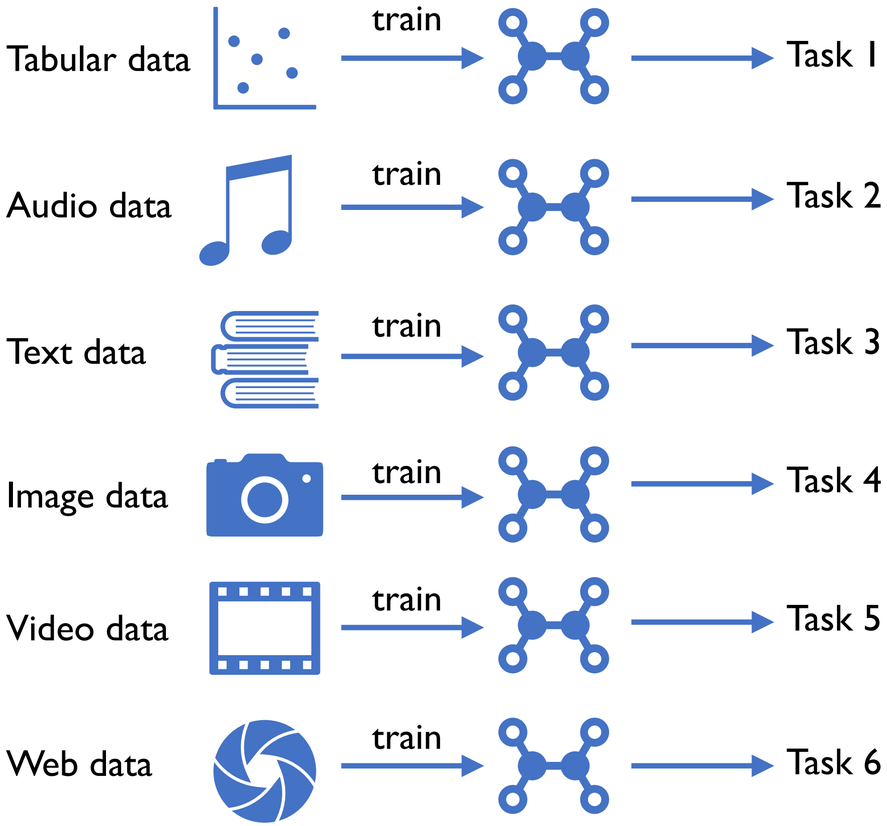
- Deep Learning enables the learning high-level features by homogenizing neural architectures without feature engineering for different applications
- ***Foundation Models*** can learn various ***functionalities*** (emergence) without training ***multiple models*** (homogenization) at scale



Bommasani et al., On the Opportunities and Risks of Foundation Models, *CRFM Stanford HAI Reports* (2022)

Traditional ML

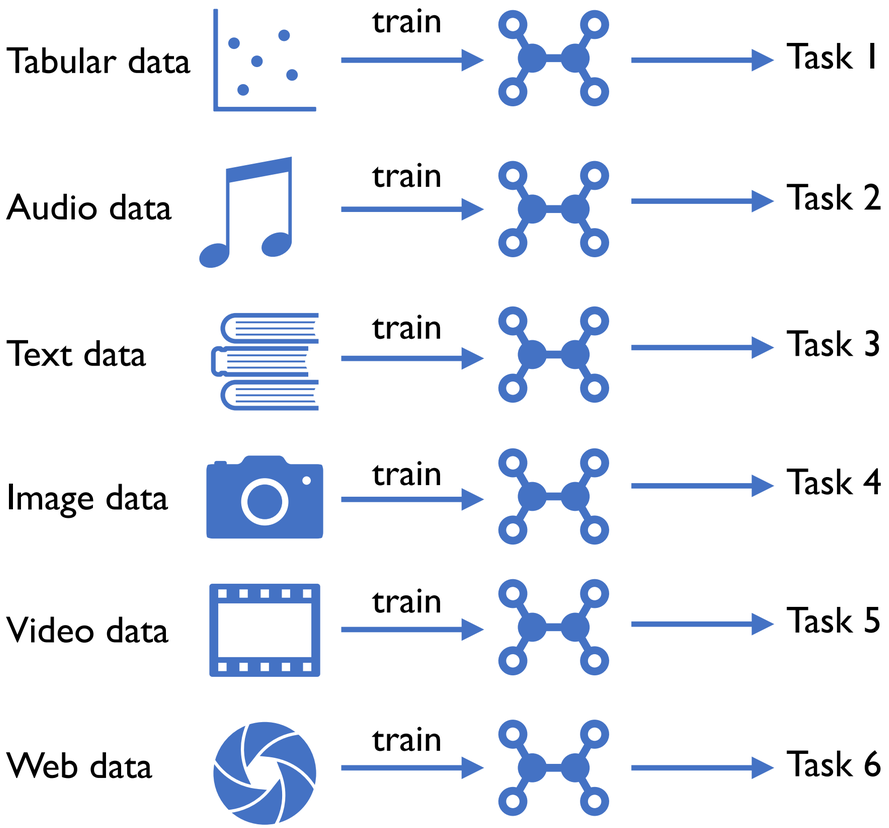
individual siloed models
require task-specific training

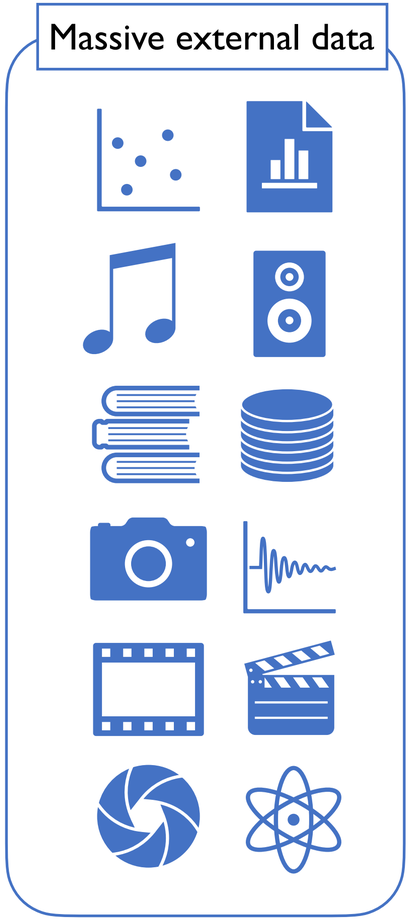


Traditional ML vs Foundation Models

individual siloed models
require task-specific training

adaptable with zero-shot or few-shot
require pre-training





Some people believe LLMs can do this but many AI scientists disagree

train

Foundation Model

Multi-tasks

- Text generation
- Question & Answering
- Information extraction
- Image captioning
- Code generation
- ⋮

Adapted Model

Additional tasks

- Classification
- Scientific reasoning
- ⋮

Formulation of Foundation Model

- **Pretraining**: pretrain a model with large-scale unlabeled dataset

unlabeled dataset $\mathcal{D}_{\text{train}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

pretraining model (foundation model) $F_\theta : \mathbf{x} \mapsto F_\theta(\mathbf{x}) \in \mathbb{R}^m$

pretraining loss $\mathcal{L}_{\text{pre}}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{pre}}(F_\theta(\mathbf{x}_i), \mathbf{x}_i)$

pretrained model $\theta_* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{\text{pre}}(\theta)$

Formulation of Foundation Model

- **Pretraining**: pretrain a model with large-scale unlabeled dataset
- **Adaptation**: adapt the pretrained model to a wide range of tasks

unlabeled dataset $\mathcal{D}_{\text{train}} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

pretraining model (foundation model) $F_\theta : \mathbf{x} \mapsto F_\theta(\mathbf{x}) \in \mathbb{R}^m$

pretraining loss $\mathcal{L}_{\text{pre}}(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell_{\text{pre}}(F_\theta(\mathbf{x}_i), \mathbf{x}_i)$

pretrained model $\theta_* = \underset{\theta}{\operatorname{argmin}}\, \mathcal{L}_{\text{pre}}(\theta)$

labeled downstream task

$$\{(\mathbf{x}_i^{(\tau)}, y_i^{(\tau)}) : i = 1, \ldots, |\mathcal{D}_\tau|\}_{\tau \in \mathcal{T}}$$

number of data downstream task

task

linear probing

$$\omega_*^{(\tau)} = \underset{\omega}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_\tau|} \sum_{i=1}^{|\mathcal{D}_\tau|} \ell_{\text{task}}(\omega^\top F_{\theta_*}(\mathbf{x}_i^{(\tau)}), y_i^{(\tau)})$$

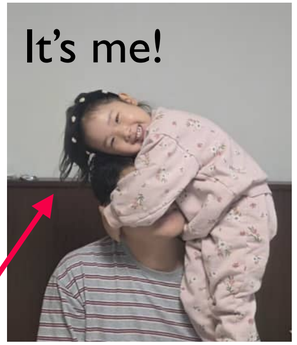
fixed at θ_*

fine tuning

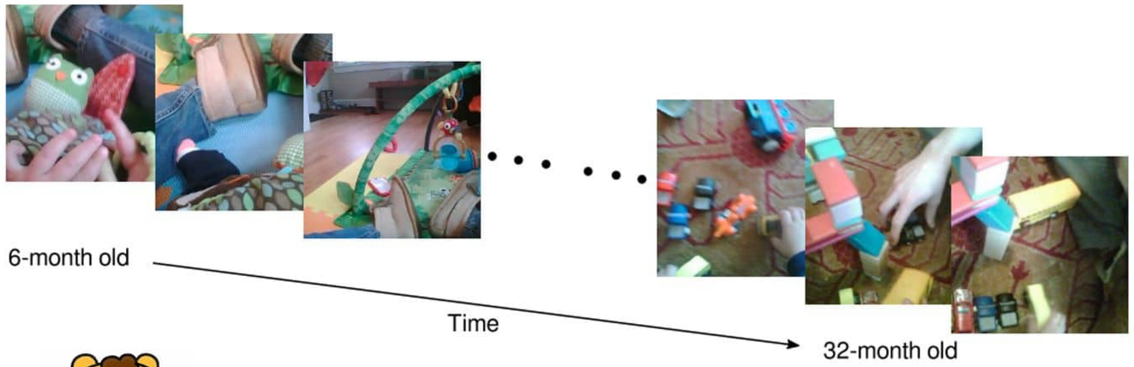
$$\omega_*^{(\tau)}, \theta_*^{(\tau)} = \underset{\omega, \theta}{\operatorname{argmin}} \frac{1}{|\mathcal{D}_\tau|} \sum_{i=1}^{|\mathcal{D}_\tau|} \ell_{\text{task}}(\omega^\top F_{\theta}(\mathbf{x}_i^{(\tau)}), y_i^{(\tau)})$$

initialized at θ_*

Self-Supervised Learning

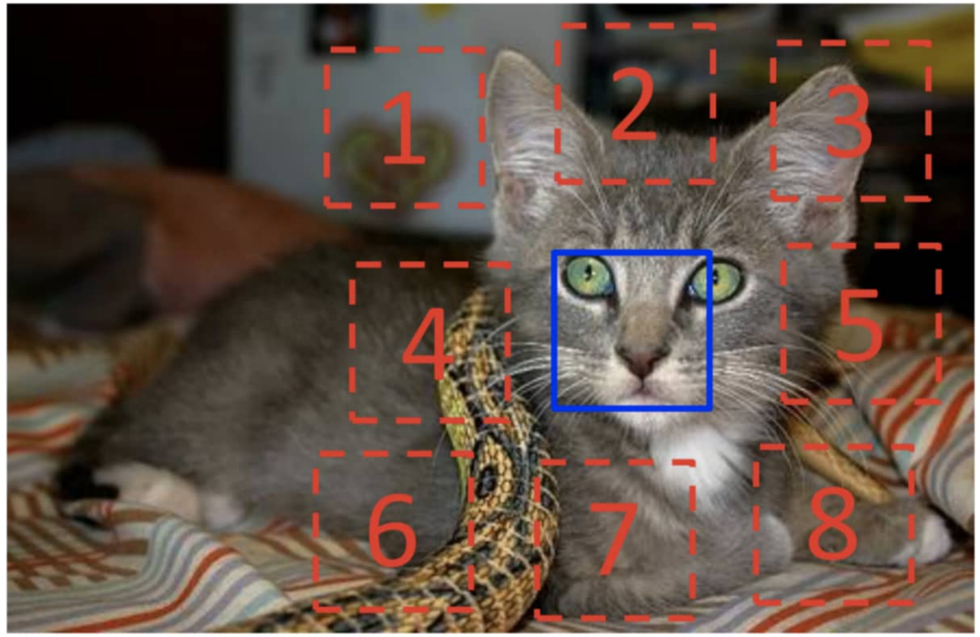


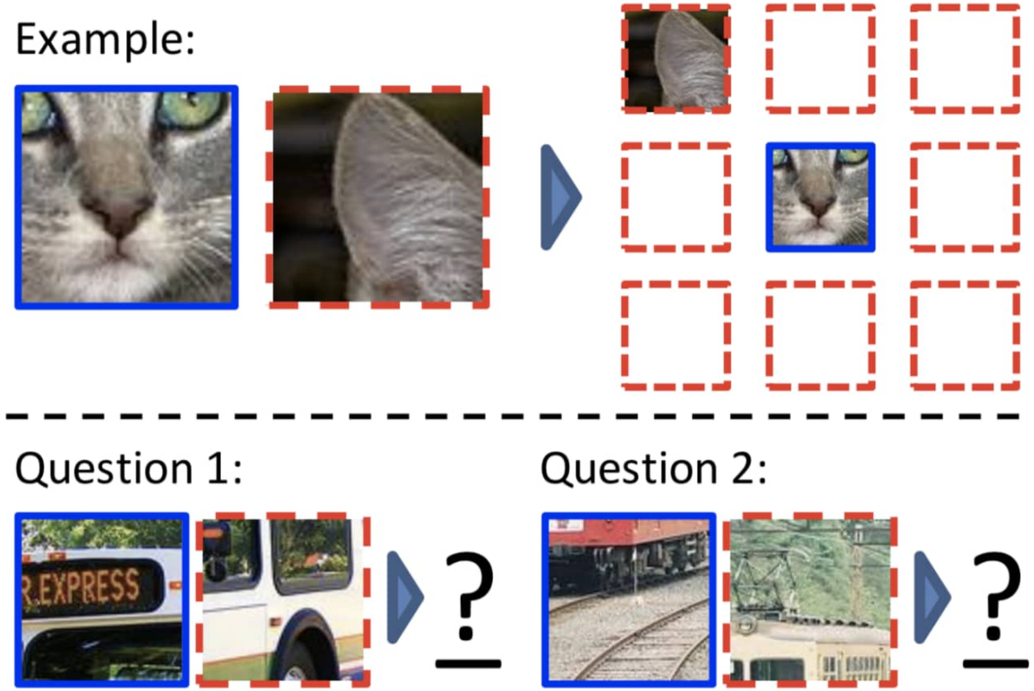
- Self-supervised learning is a form of ***unsupervised training*** where the data itself provides the supervision signal
  - the representations learned on the pretext task are subsequently used for a different downstream task



Children can learn high-level visual representations without strong supervision

Pretraining Example: Position Prediction





Doersch et al., Unsupervised Visual Representation Learning by Context Prediction, *ICCV* (2015)

Pretraining Example: Contrastive Learning

pull the similar sample pairs

$$\ell_{\text{triplet}}(x, x^+) := \|f_\theta(x) - f_\theta(x^+)\|$$

anchor x^a

x^+ positive sample

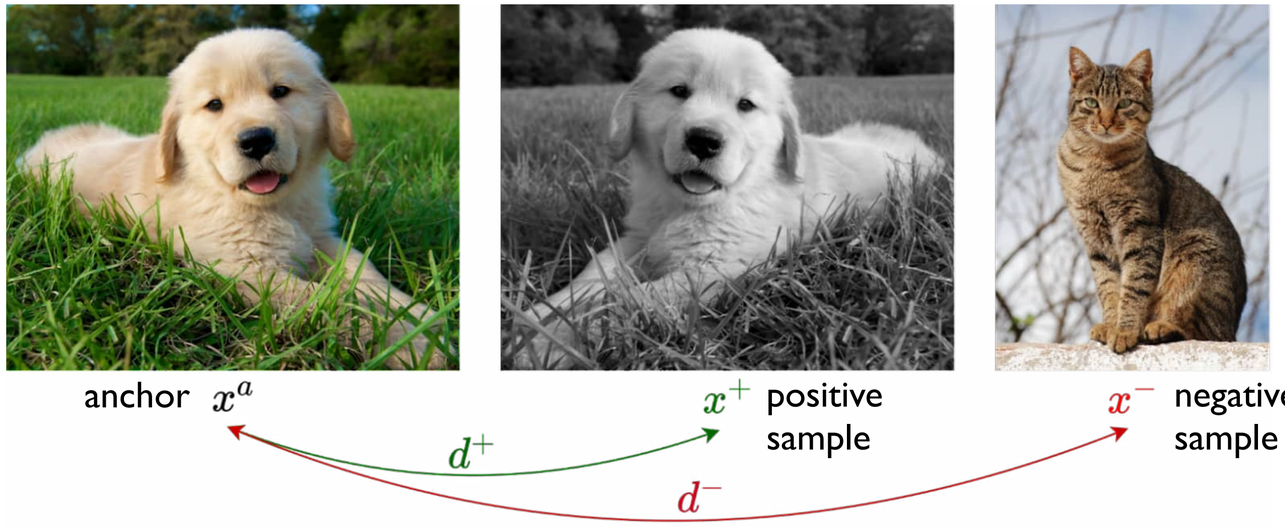
Schroff et al., FaceNet: A Unified Embedding for Face Recognition and Clustering, ***CVPR*** (2015)

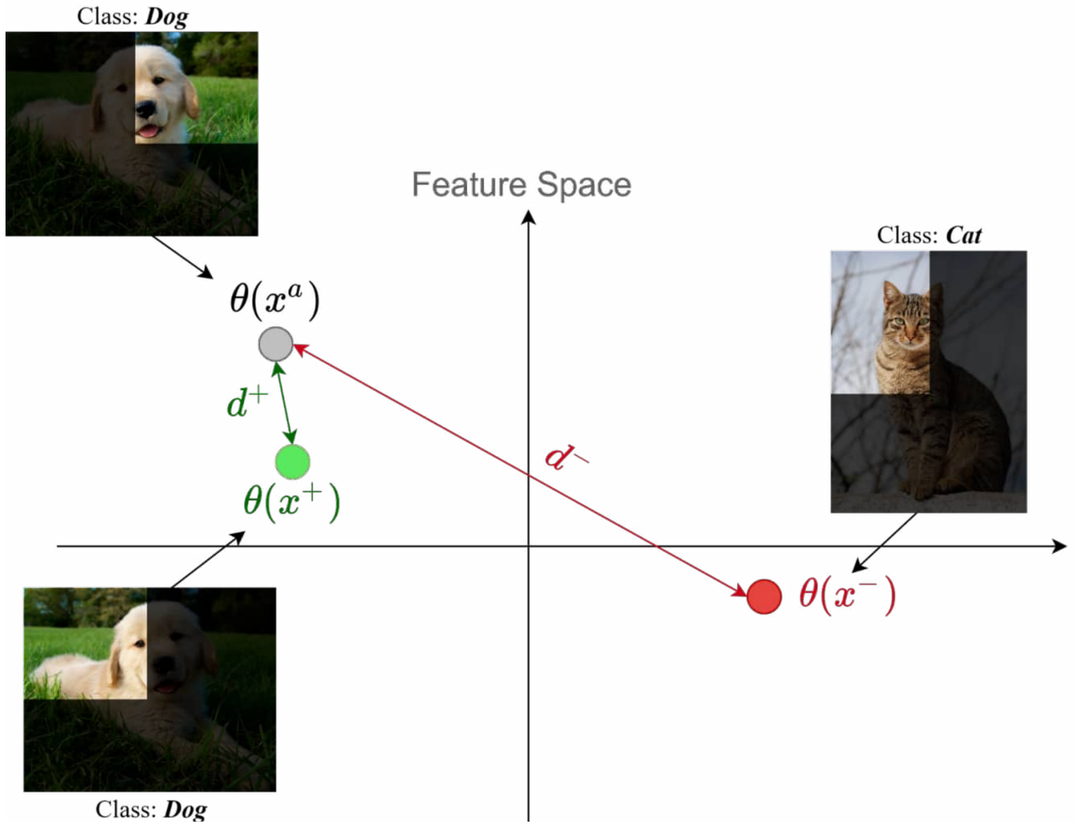
Pretraining Example: Contrastive Learning

pull the similar sample pairs　　push away negative pairs

$$\ell_{\text{triplet}}(x, x^+, x^-) := \|f_\theta(x) - f_\theta(x^+)\| - \|f_\theta(x) - f_\theta(x^-)\|$$

triplet loss





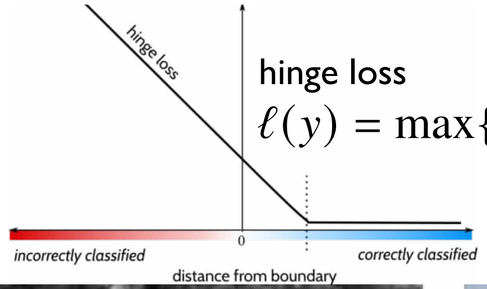
Schroff et al., FaceNet: A Unified Embedding for Face Recognition and Clustering, ***CVPR*** (2015)

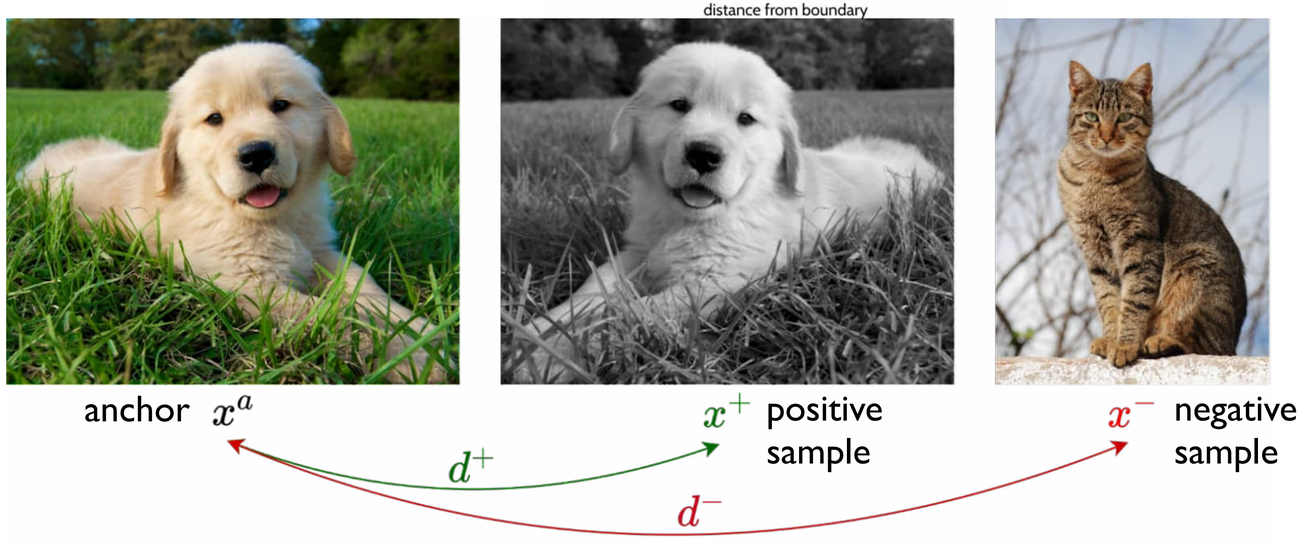
Pretraining Example: Contrastive Learning

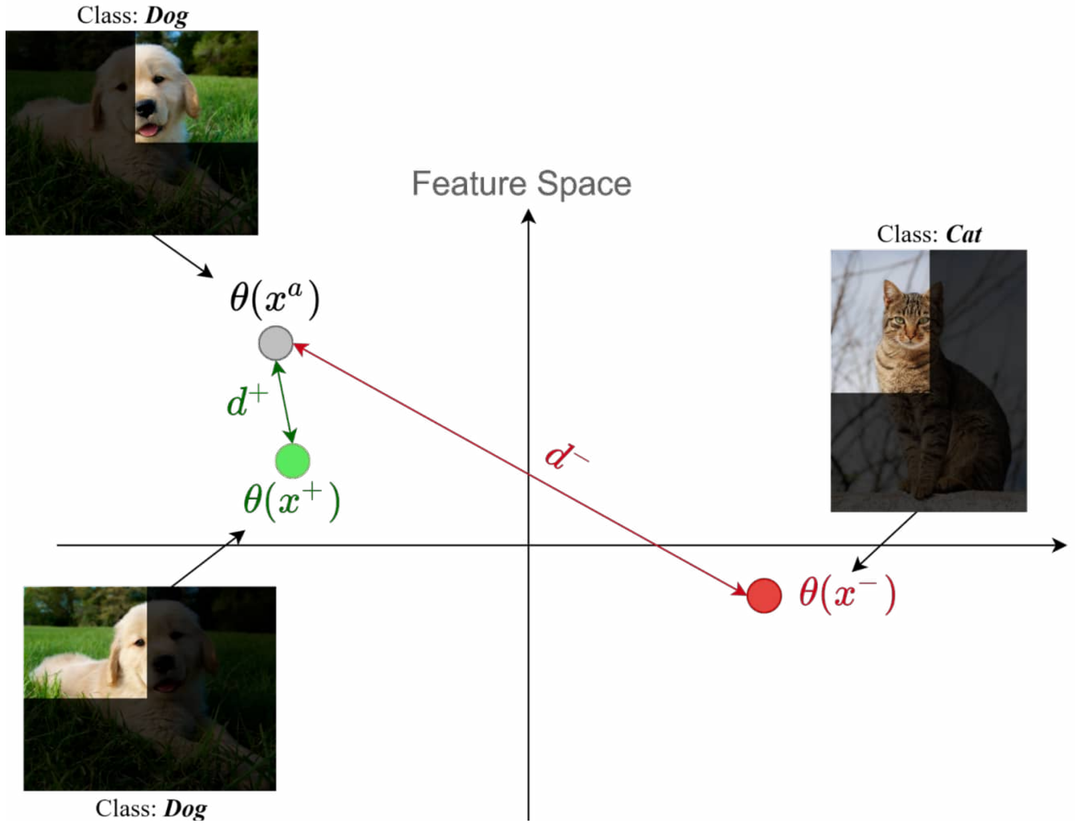
bound — pull the similar sample pairs — push away negative pairs — margin

$$\ell_{\text{triplet}}(x, x^+, x^-) := \max\left\{\|f_\theta(x) - f_\theta(x^+)\| - \|f_\theta(x) - f_\theta(x^-)\| + \epsilon, 0\right\}$$

triplet loss

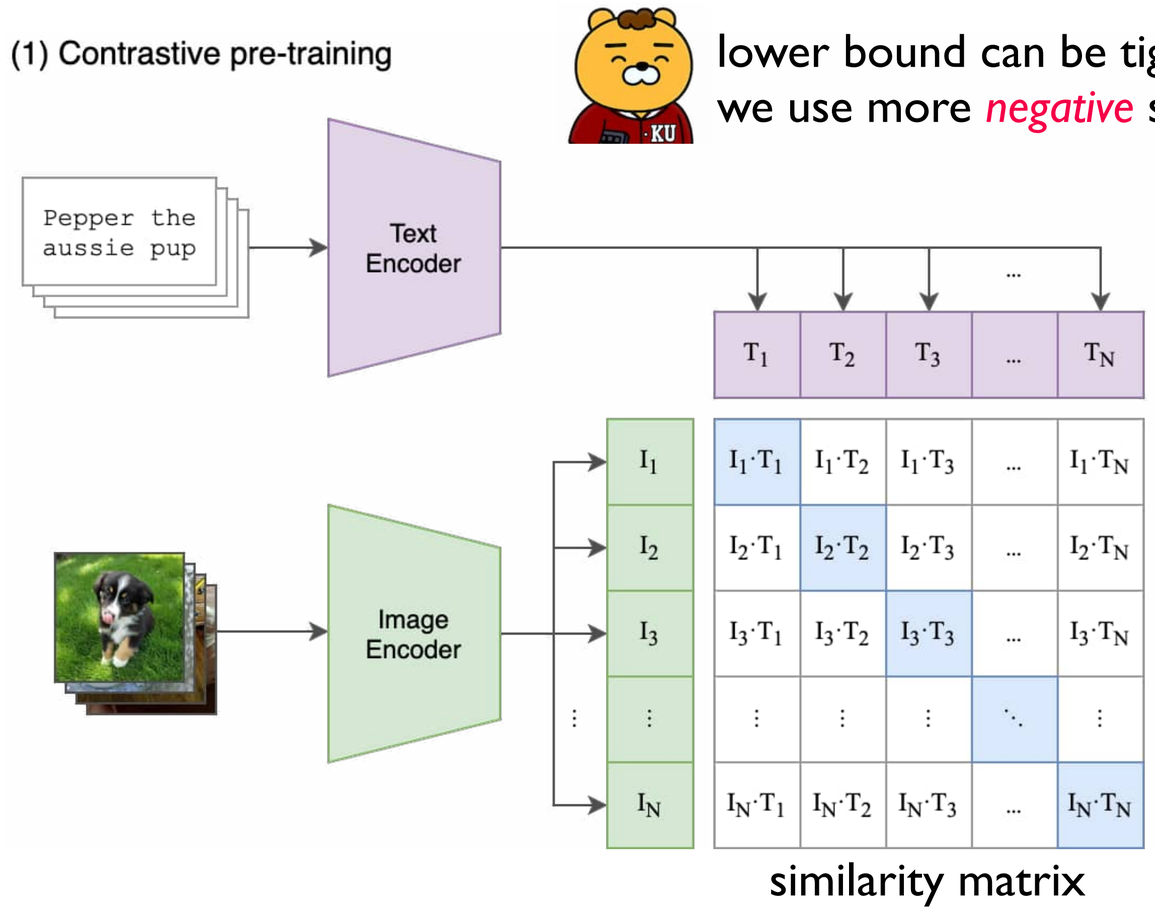






Schroff et al., FaceNet: A Unified Embedding for Face Recognition and Clustering, ***CVPR*** (2015)

Contrastive Learning as Density Ratio Estimation



lower bound can be tighter as we use more *negative* samples

$$\mathbb{KL}(P(I,T)\|P(I)P(T)) \geq \log N - L_{\text{InfoNCE}}$$

mutual information

InfoNCE (Noise-Contrastive Estimation)

$$\text{img2txt} \quad -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp(I_n \cdot T_n)}{\sum_{m=1}^{N}\exp(I_n \cdot T_m)}$$

$$\text{txt2img} \quad -\frac{1}{N}\sum_{m=1}^{N}\log\frac{\exp(I_m \cdot T_m)}{\sum_{n=1}^{N}\exp(I_n \cdot T_m)}$$

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, *ICML* (2021)

How can computers *generate* image and text?

Do they draw, write, and dance?

Everything is numbers on computer!

image

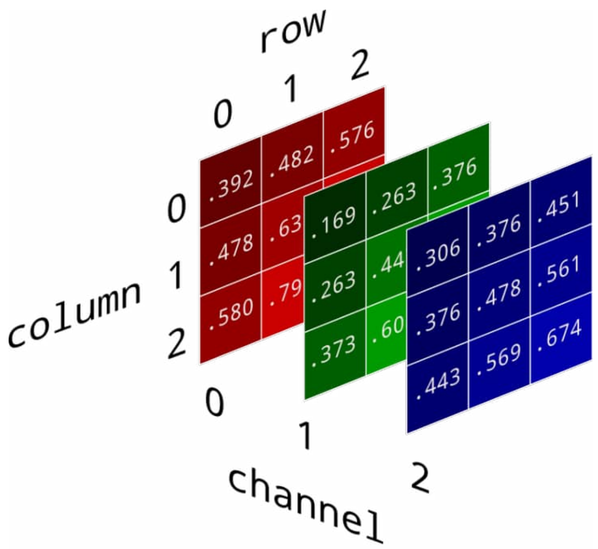
I Love
Korea Univ.

natural language

Everything is numbers on computer!

image

RGB



Recognition

binary

computers

I Love
Korea Univ.

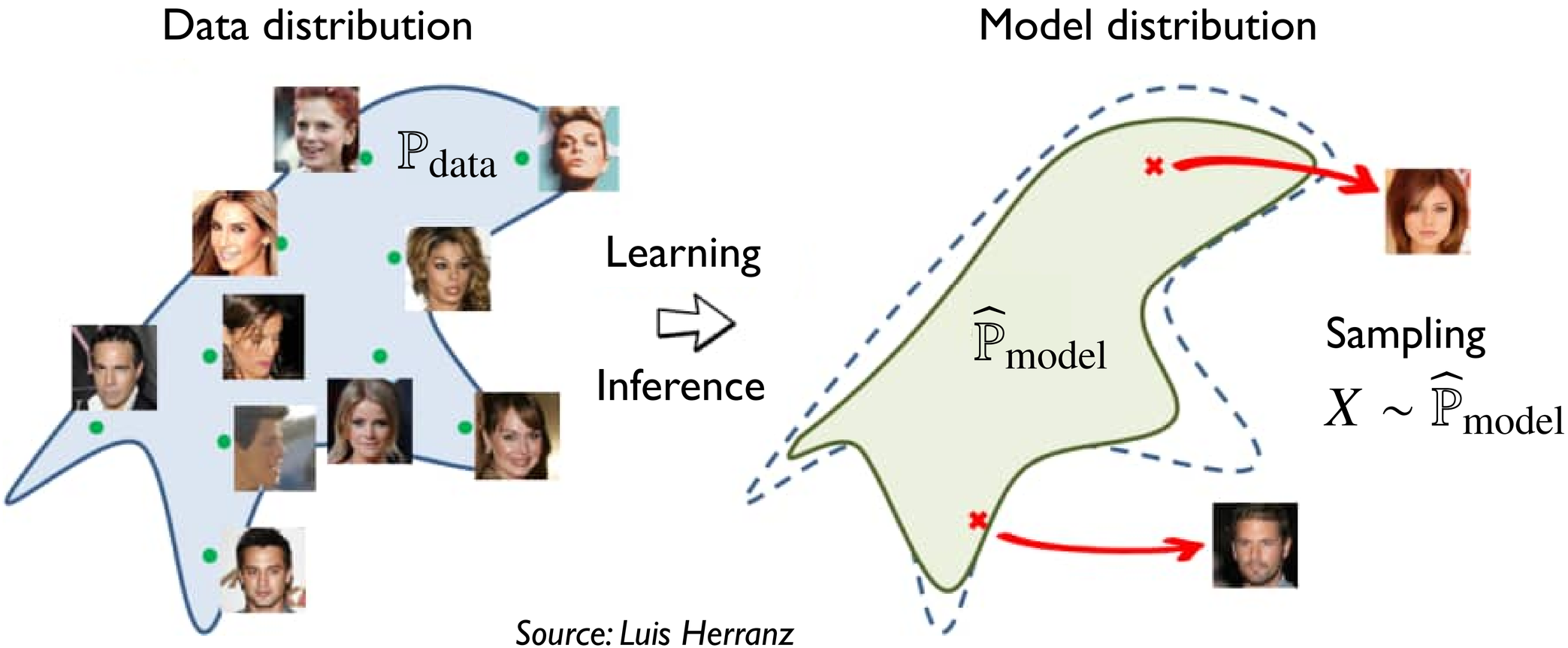
natural language

Unicode

Dec	Symbol	Binary	Dec	Symbol	Binary
65	A	0100 0001	83	S	0101 0011
66	B	0100 0010	84	T	0101 0100
67	C	0100 0011	85	U	0101 0101
68	D	0100 0100	86	V	0101 0110
69	E	0100 0101	87	W	0101 0111
70	F	0100 0110	88	X	0101 1000
71	G	0100 0111	89	Y	0101 1001
72	H	0100 1000	90	Z	0101 1010
73	I	0100 1001	91	[	0101 1011
74	J	0100 1010	92	\	0101 1100
75	K	0100 1011	93	]	0101 1101
76	L	0100 1100	94	^	0101 1110
77	M	0100 1101	95	_	0101 1111
78	N	0100 1110	96	`	0110 0000
79	O	0100 1111	97	a	0110 0001
80	P	0101 0000	98	b	0110 0010
81	Q	0101 0001	99	c	0110 0011
82	R	0101 0010	100	d	0110 0100

Generation

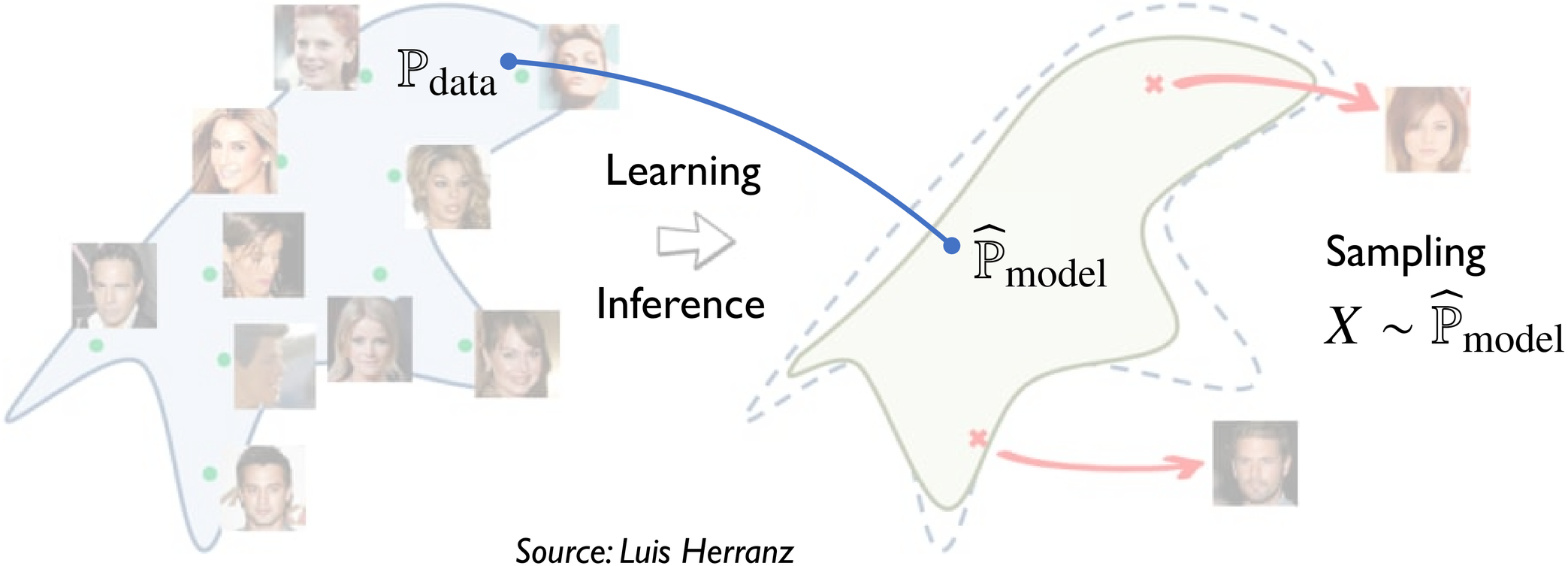
Principles of Generative AI



Source: Luis Herranz

Principles of Generative AI

Data distribution $\mathbb{P}_{\text{data}} \approx \widehat{\mathbb{P}}_{\text{model}}$ Model distribution



Source: Luis Herranz

Statistical Principles of Generative Model

$$\mathrm{dist}(\mathbb{P}_{\text{data}}, \widehat{\mathbb{P}}_{\text{model}}) \searrow 0$$

Statistical Principles of Generative Model

Kullback-Leibler Divergence

$$\mathbb{KL}(\mathbb{P}_{\text{data}} \| \widehat{\mathbb{P}}_{\text{model}}) \searrow 0$$

$$= \mathbb{E}_{X \sim P_{\text{data}}(x)} \left[\log \frac{P_{\text{data}}(X)}{\widehat{P}_{\text{model}}(X)} \right]$$

Statistical Principles of Generative Model

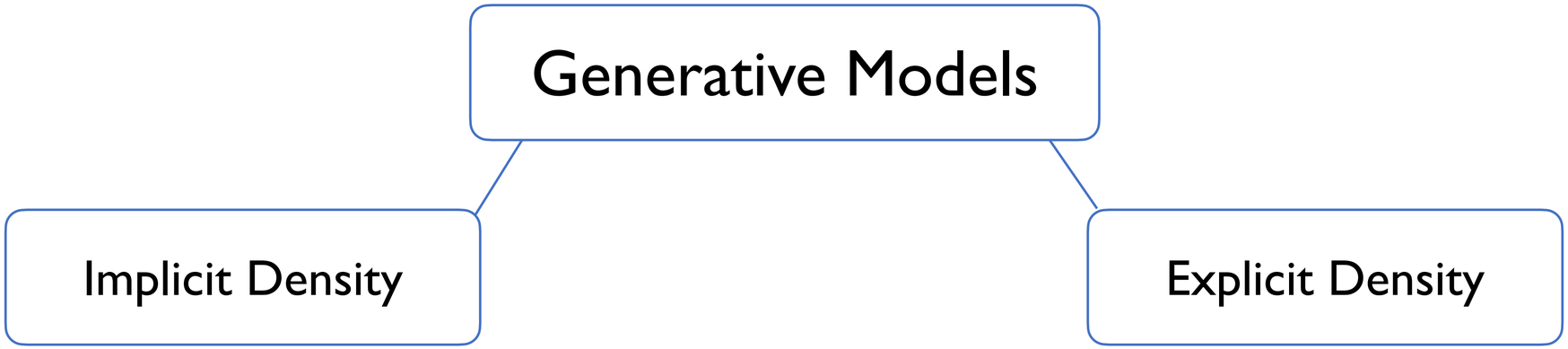
$$\widehat{P}_* = \arg\min_{\widehat{P}_{\text{model}}} \mathbb{KL}(\mathbb{P}_{\text{data}} \| \widehat{\mathbb{P}}_{\text{model}})$$

$$= \arg\min_{\widehat{P}_{\text{model}}} \mathbb{E}_{X \sim P_{\text{data}}} \left[\log P_{\text{data}}(X) - \log \widehat{P}_{\text{model}}(X) \right]$$

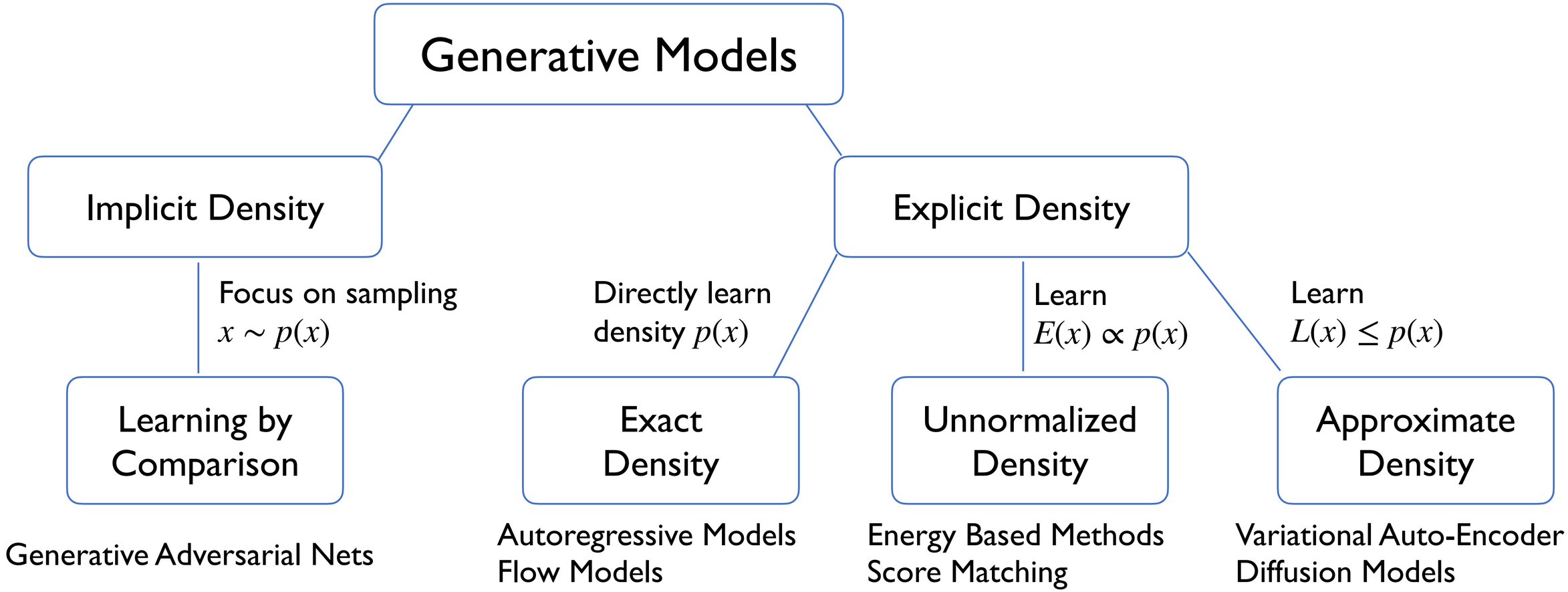
$$= \arg\max_{\widehat{P}_{\text{model}}} \mathbb{E}_{X \sim P_{\text{data}}} \left[\log \widehat{P}_{\text{model}}(X) \right] \Rightarrow \textit{Maximum Likelihood!}$$

Maximum Likelihood Principle is an essential key to understand *Generative Models*!

Taxonomy of Generative Models



Taxonomy of Generative Models



Taxonomy of Generative Models

modern generative AIs focus on
probabilistic generative modeling

- **Generative Models**
  - **Explicit Density**
    - Directly learn density $p(x)$ → **Exact Density**
      - Autoregressive Models
      - Flow Models
    - Learn $E(x) \propto p(x)$ → **Unnormalized Density**
      - Energy Based Methods
      - Score Matching
    - Learn $L(x) \leq p(x)$ → **Approximate Density**
      - Variational Auto-Encoder
      - Diffusion Models
  - **Implicit Density**
    - Focus on sampling $x \sim p(x)$
    - **Learning by Comparison**
      - Generative Adversarial Nets

Beyond Probabilistic Reasoning

Toward Causal Reasoning (CaR)

Probabilistic Reasoning

EVI

MAR

CON

MAP

Cause-Effect Analysis

Interventional Reasoning

Counterfactual Reasoning

Causal Reasoning

	Layer (Symbolic)	Typical Activity	Typical Question	Example	Machine Learning
$\mathcal{L}_1$	Associational $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell us about the disease?	Supervised / Unsupervised Learning
$\mathcal{L}_2$	Interventional $P(y\|do(x),c)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured?	Reinforcement Learning
$\mathcal{L}_3$	Counterfactual $P(y_x\|x',y')$	Imagining	Why? What if I had acted differently?	Was it the aspirin that stopped my headache?	

Source: Judea Pearl

Understanding Causality by Generative Models

- **Structural Causal Models (SCMs) are data generative processes** describing the causal relationships between variables with causal queries (Pearl, 2009)
- Causal discovery from observational data is at the core of causality since the causal graph support the prediction of the queries (Schölkopf et al., 2021)
- Causal queries can be answered via **conditional latent variable models** which learned a proxy for the noise and structural equations (Chao et al., 2024)
- Causal reasoning can be enhanced via **functional diffusion models** which learned structural relationships in causal graphs (Kang et al., 2025)

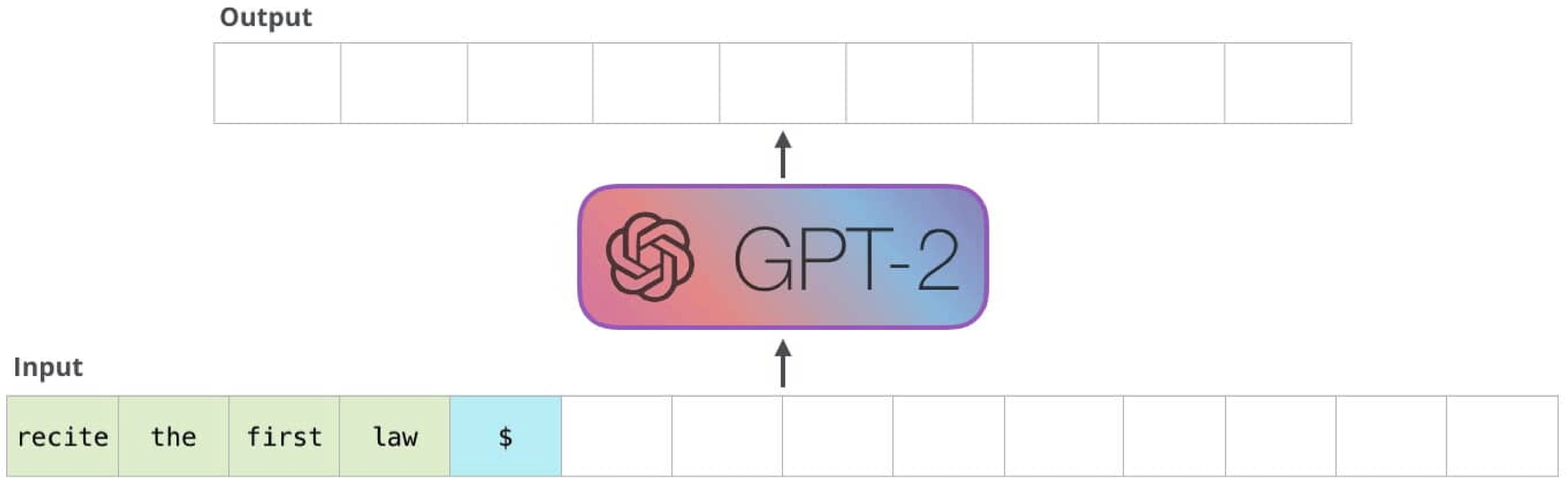
J. Pearl, Causal inference in statistics: An overview. ***Statistics Surveys,* (2009)**
B. Schölkopf et al., Toward Causal Representation Learning, ***Proceedings of the IEEE* (2021)**
P. Chao et al., Modeling Causal Mechanisms with Diffusion Models for Interventional and Counterfactual Queries, ***TMLR* (2024)**
Kang et al., Score-informed Neural Operator for Enhancing Ordering-based Causal Discovery, ***NeurIPS* (2025)**

Part 3

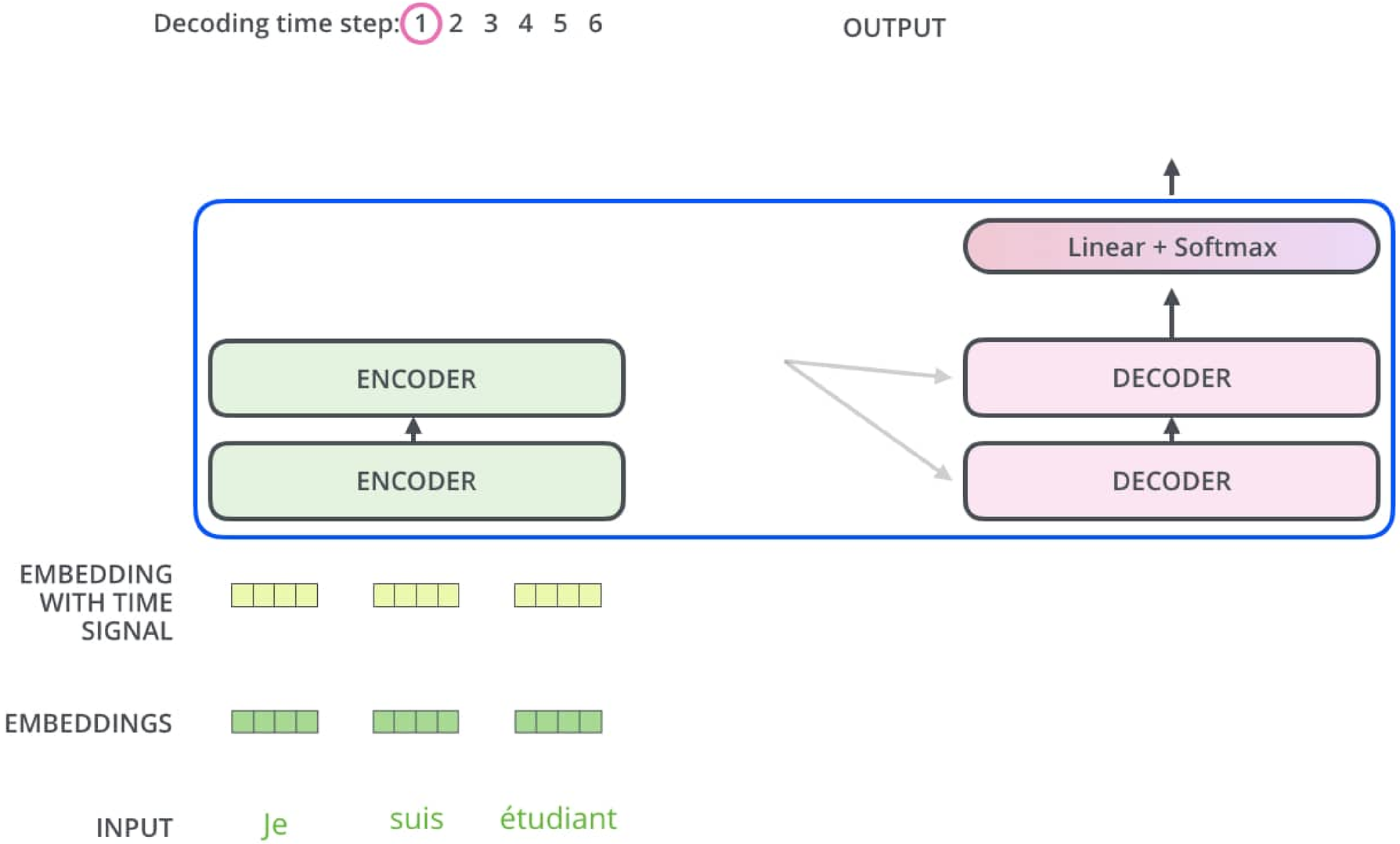
In-Context Learning & Prior-fitted

Generative Pretraining

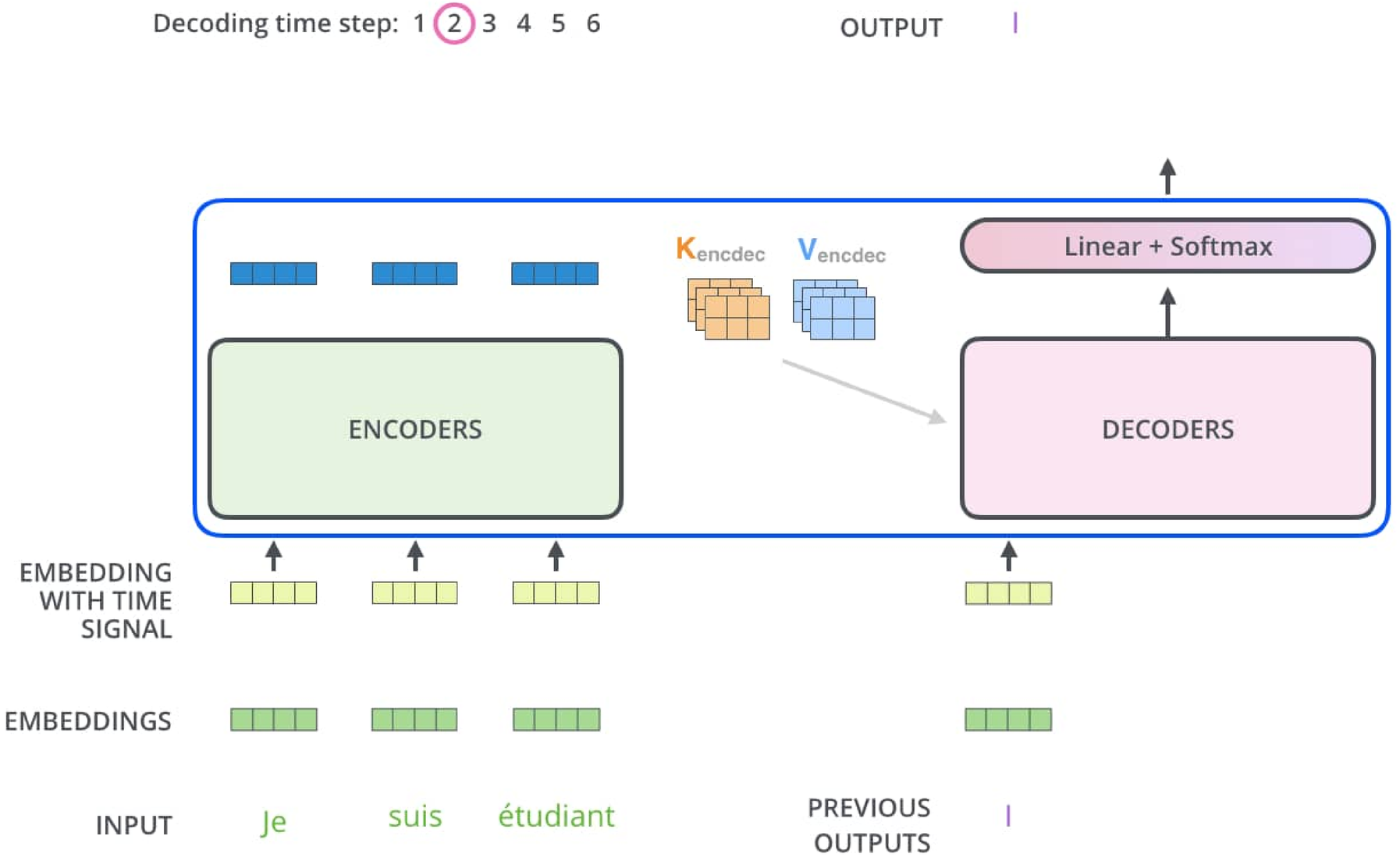
Autoregressive Generative Model $P(X) = \prod P(\mathbf{x}_t | \mathbf{x}_{<t})$



Source: Jay Alammar



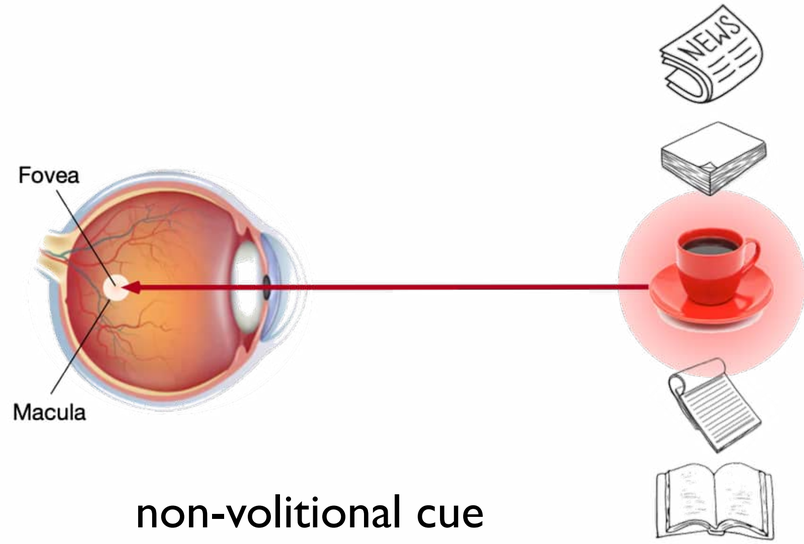
Source: Jay Alammar



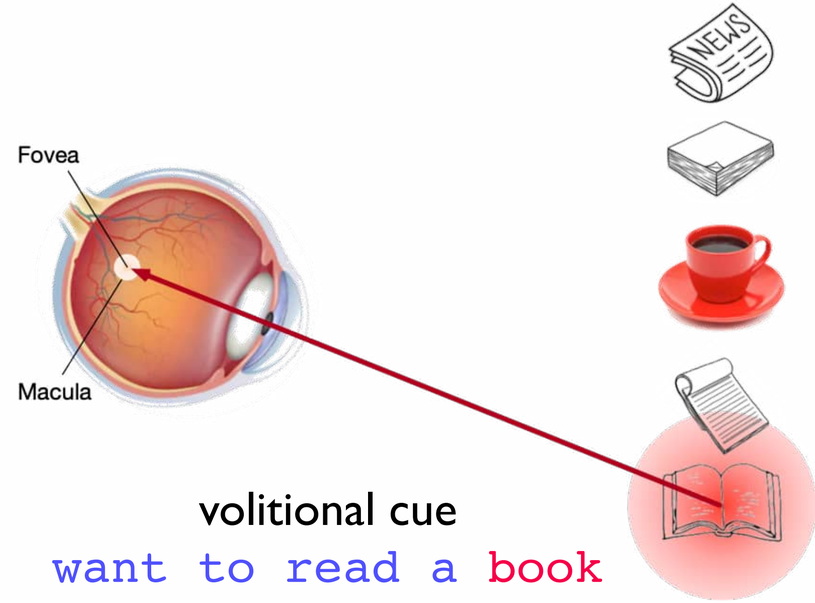
Source: Jay Alammar

Attention Module for Context Processing

- Attention enable the human to prioritize the perception in order to deal effectively with others



non-volitional cue

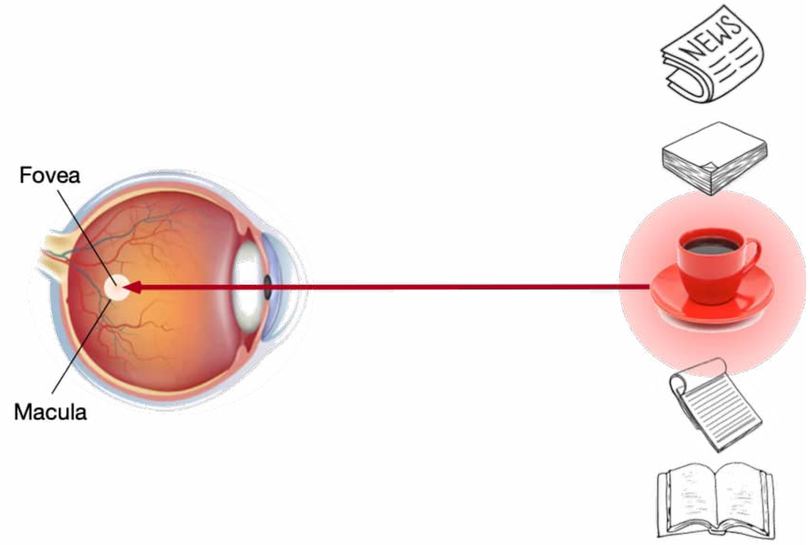


volitional cue

want to read a book

How can we formulate Attention?

list of keys

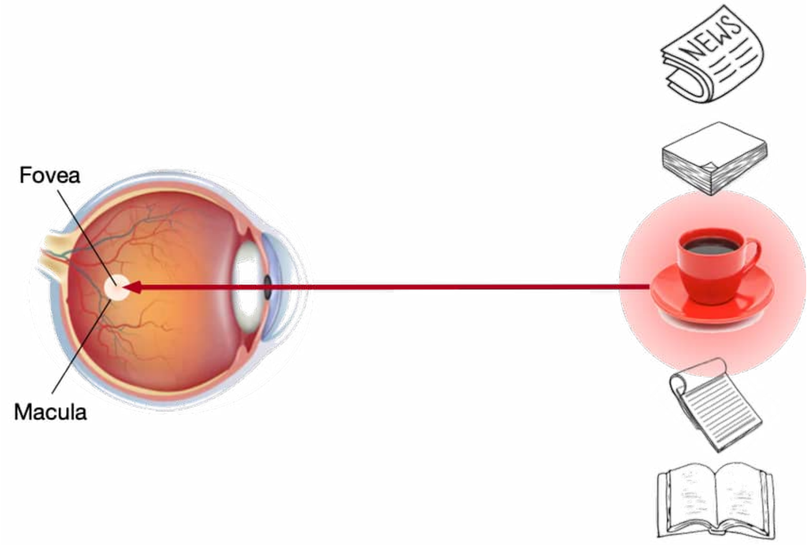


$$\mathbf{k}_i$$

key index of key

key is an item that can be *memorized* so that we can *retrieve* it from the *list of keys*

How can we formulate Attention?



list of keys

list of values

$$g_\phi(\mathbf{k}_1)$$

$$g_\phi(\mathbf{k}_2)$$

$$\vdots$$

$$g_\phi(\mathbf{k}_i)$$

i-th value

$$g_\phi(\mathbf{k}_i)$$

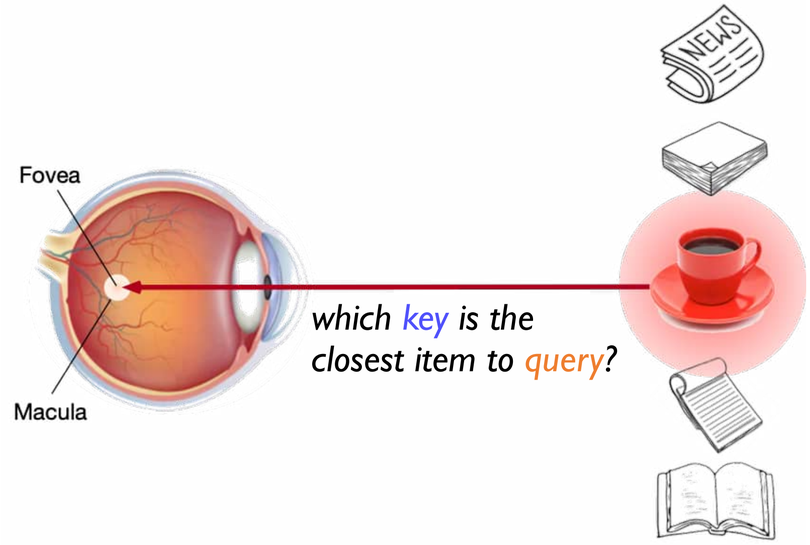
value parameter

key

index of key

a value can be an assigned output
or a feature representation for each key

How can we formulate Attention?



list of keys

list of values

$g_\phi(\mathbf{k}_1)$

$g_\phi(\mathbf{k}_2)$

$\vdots$

$g_\phi(\mathbf{k}_i)$

volitional cue

want to read a book

query $\mathbf{q}$

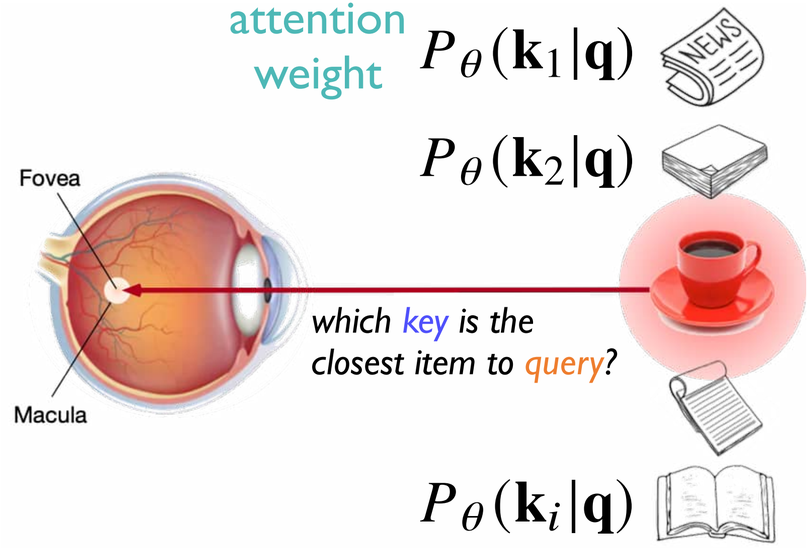
i-th value

$$g_\phi(\mathbf{k}_i)$$

value parameter — key — index of key

query is an input that is used for searching the closest key from the list of keys

How can we formulate Attention?



i-th value $g_\phi(\mathbf{k}_i)$

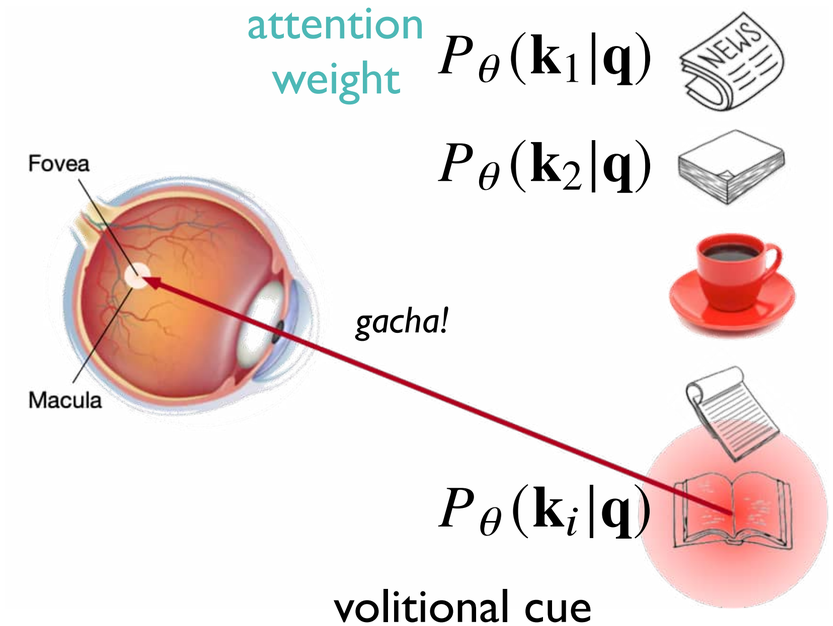
value parameter, key, index of key

i-th attention weight $\mathbb{P}_\theta(\mathbf{k}_i|\mathbf{q})$

attention parameter, query

the attention weight can be computed by representing it as the conditional probability

How can we formulate Attention?



list of keys

list of values

attention weight

$P_\theta(\mathbf{k}_1|\mathbf{q})$ $g_\phi(\mathbf{k}_1)$

$P_\theta(\mathbf{k}_2|\mathbf{q})$ $g_\phi(\mathbf{k}_2)$

$\vdots$

$P_\theta(\mathbf{k}_i|\mathbf{q})$ $g_\phi(\mathbf{k}_i)$

volitional cue

want to read a book

query $\mathbf{q}$

i-th value $g_\phi(\mathbf{k}_i)$

value parameter, key, index of key

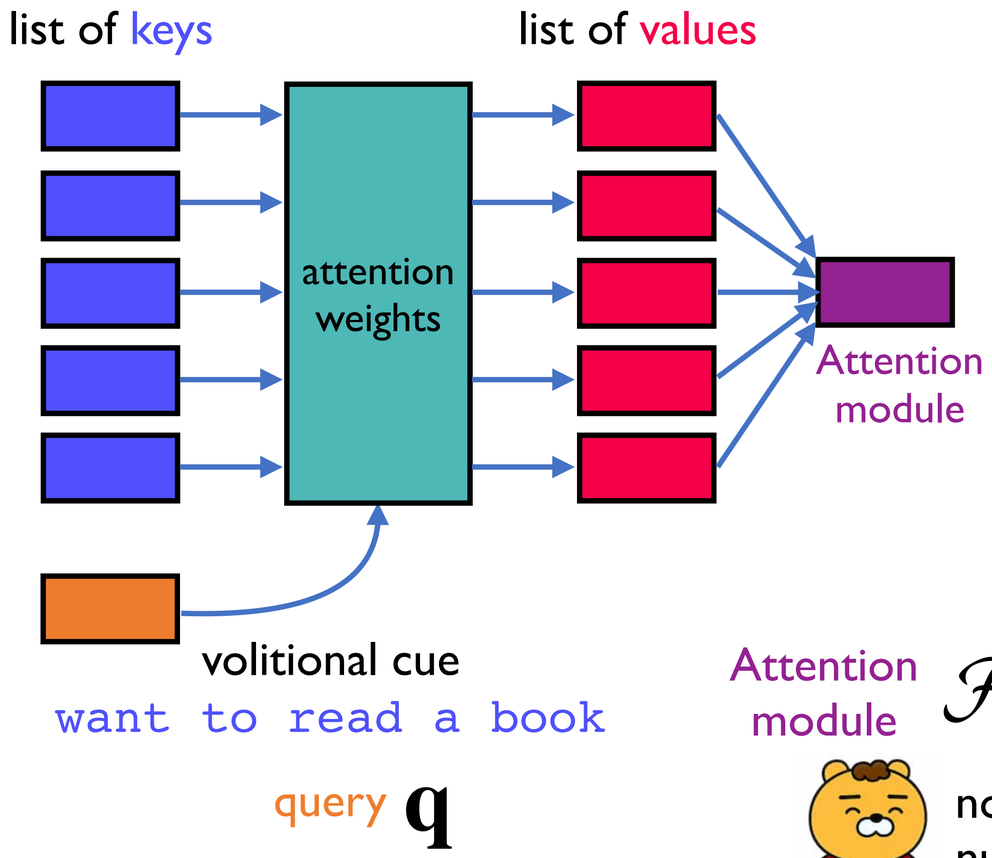
i-th attention weight $\mathbb{P}_\theta(\mathbf{k}_i|\mathbf{q})$

attention parameter, query

Attention module $\mathcal{A}_{\theta,\phi}(\mathbf{q}, \{\mathbf{k}_i\}) = \mathbb{E}_{\mathbf{k}\sim P_\theta(\cdot|\mathbf{q})}[g_\phi(\mathbf{k})]$

we can compute the conditional expectation of value with respect to the attention weights as conditional probability!

How can we formulate Attention?



i-th value $g_\phi(\mathbf{k}_i)$

value parameter, key, index of key

i-th attention weight $\mathbb{P}_\theta(\mathbf{k}_i|\mathbf{q})$

attention parameter, query

Attention module $\mathcal{A}_{\theta,\phi}(\mathbf{q}, \{\mathbf{k}_i\}) = \sum_i P_\theta(\mathbf{k}_i|\mathbf{q}) g_\phi(\mathbf{k}_i)$

attention weight, value, key

note that we are using a finite number of keys for computing!

Self-Attention Module

- We can feed a sequence of tokens into attention pooling so that the same set of tokens act as queries, keys, and values: this is called **self-attention**

$$\underbrace{\mathcal{A}_{\theta,\phi}(\mathbf{q}, \{\mathbf{k}_i\})}_{\text{general attention module}} = \left(\sum_i \frac{\exp\left(a_\theta(\mathbf{q}_t, \mathbf{k}_i)\right)}{\sum_j \exp\left(a_\theta(\mathbf{q}_t, \mathbf{k}_j)\right)} g_\phi(\mathbf{k}_i) \right)_{t=1}^{T}$$

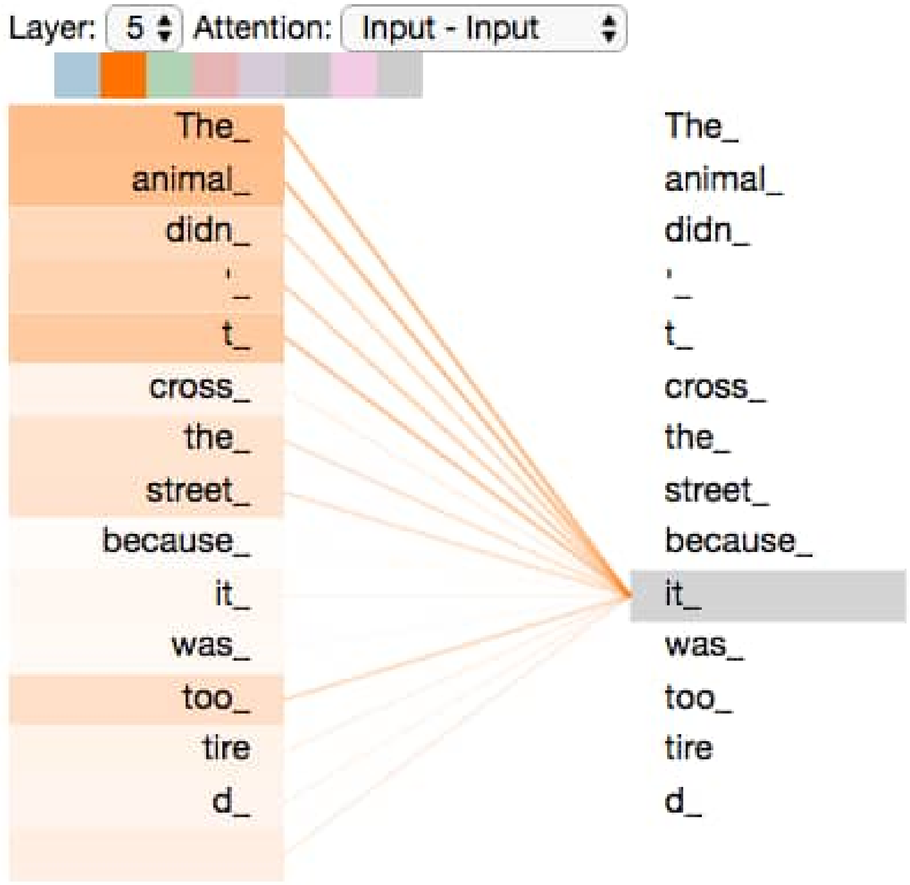
$$\mathbf{q} = (\mathbf{q}_1, \ldots, \mathbf{q}_T) \in \mathbb{R}^{T \times d_q}$$

sequence of queries

$$\mathbf{k} = \{\mathbf{k}_1, \ldots, \mathbf{k}_m\} \in \mathbb{R}^{m \times d_k}$$

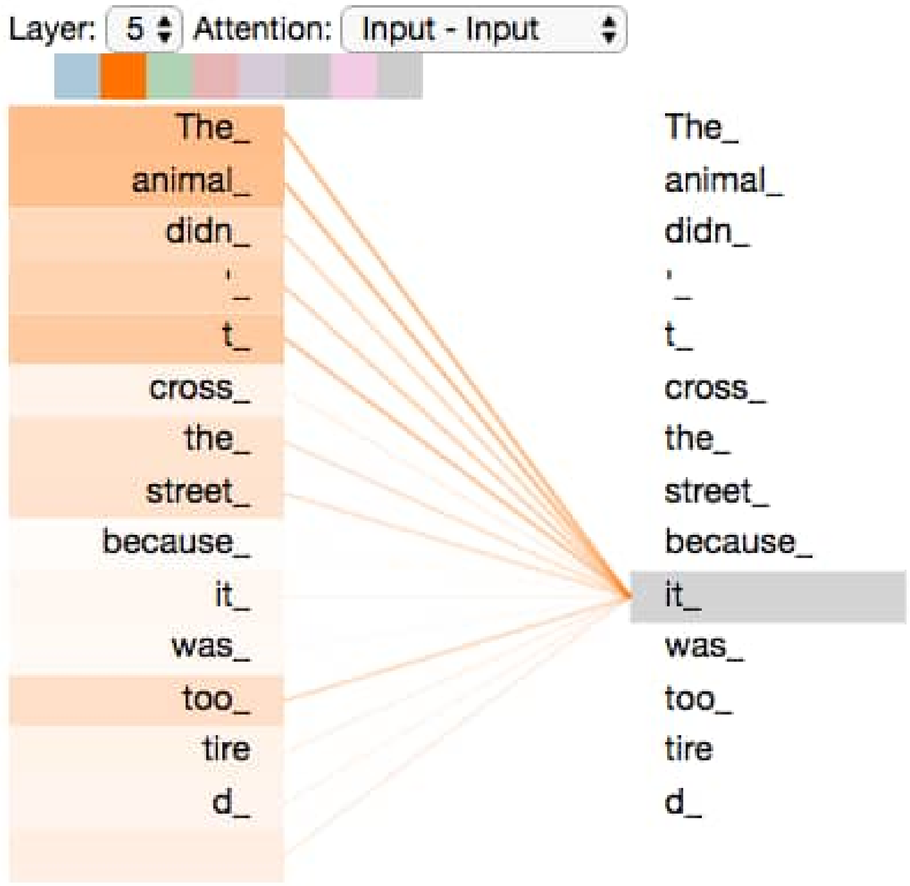
list of keys

Understanding Multi-head Self-Attention

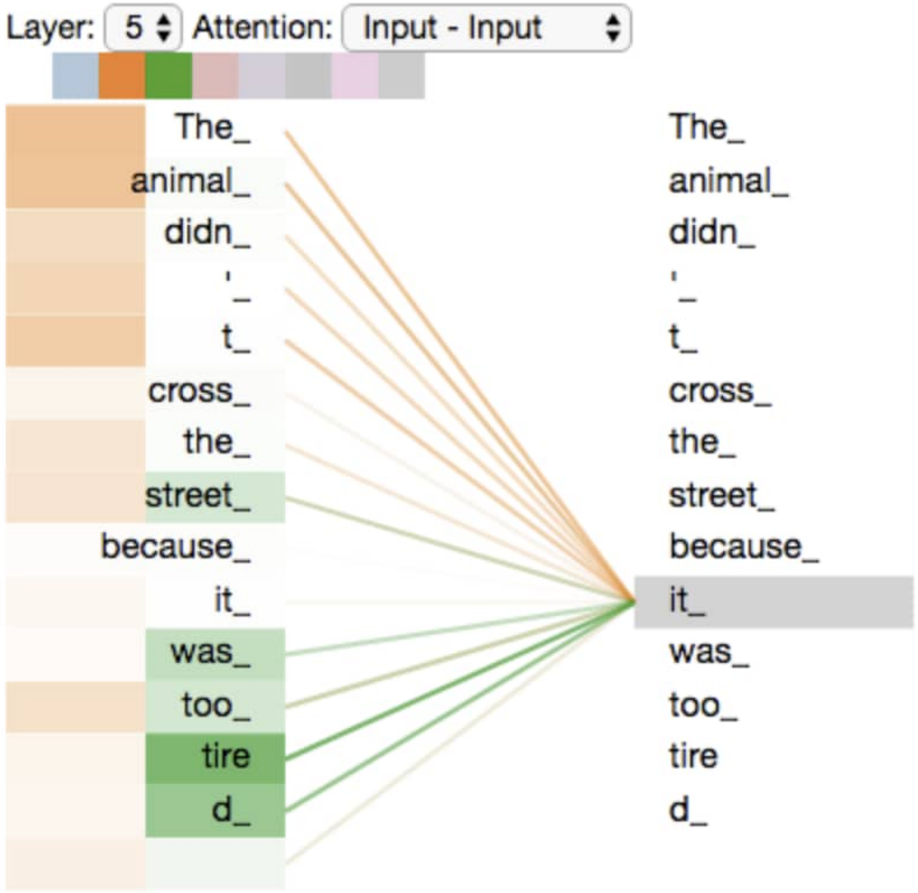


single-head self-attention

Understanding Multi-head Self-Attention

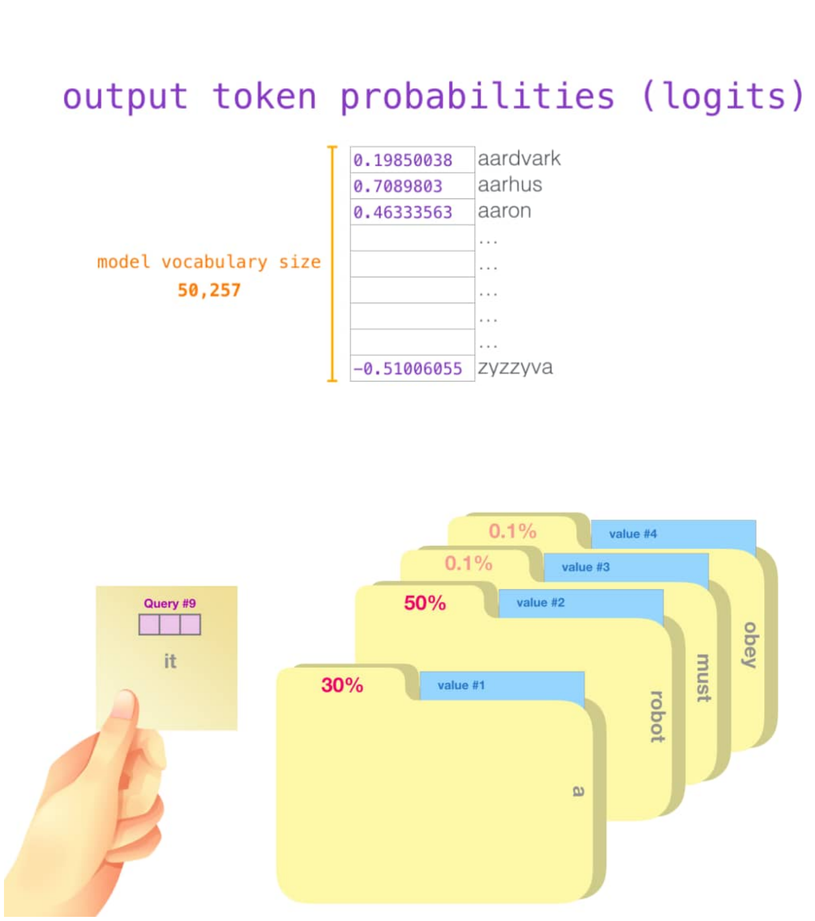


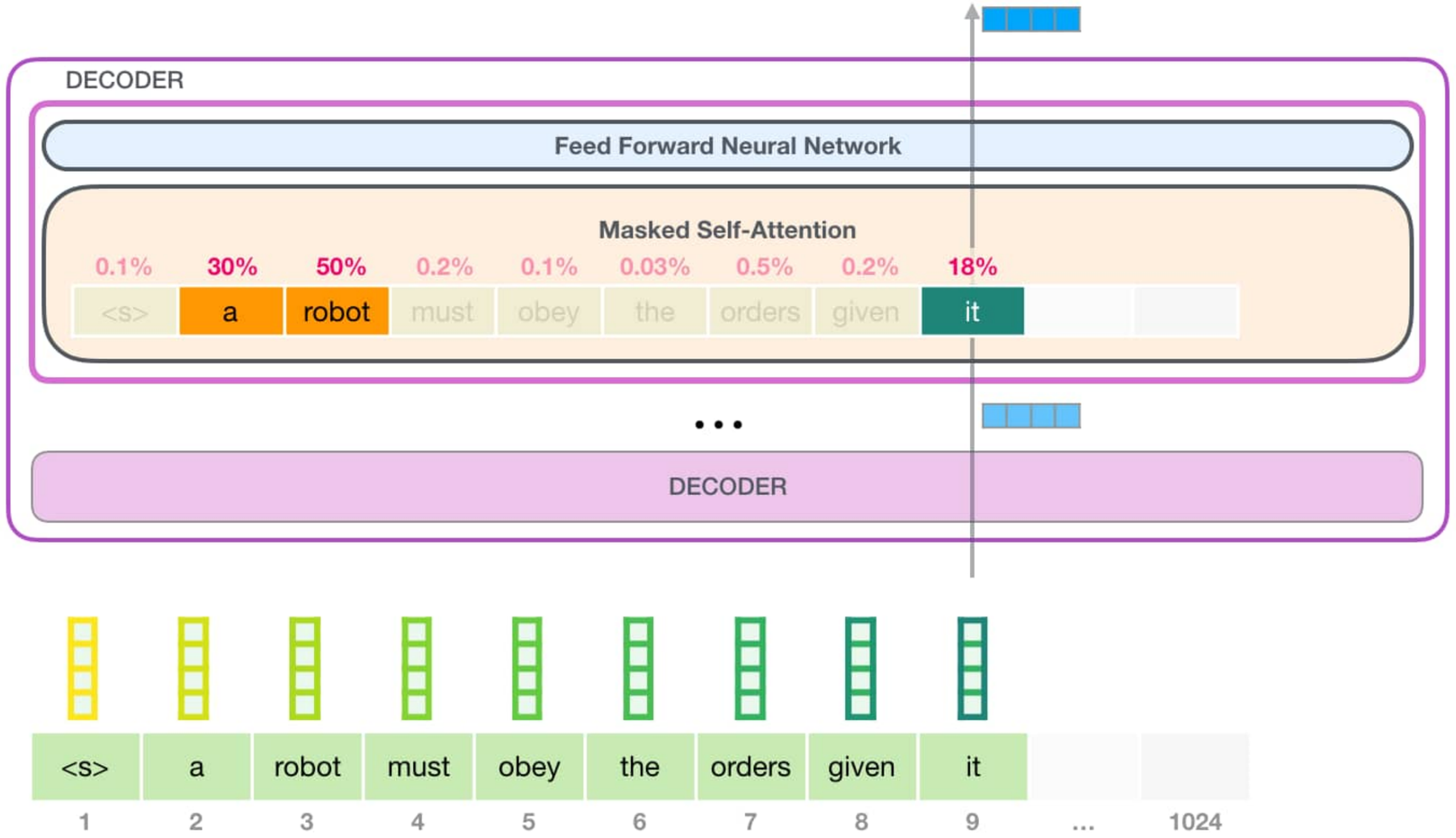
single-head self-attention



multi-head self-attention

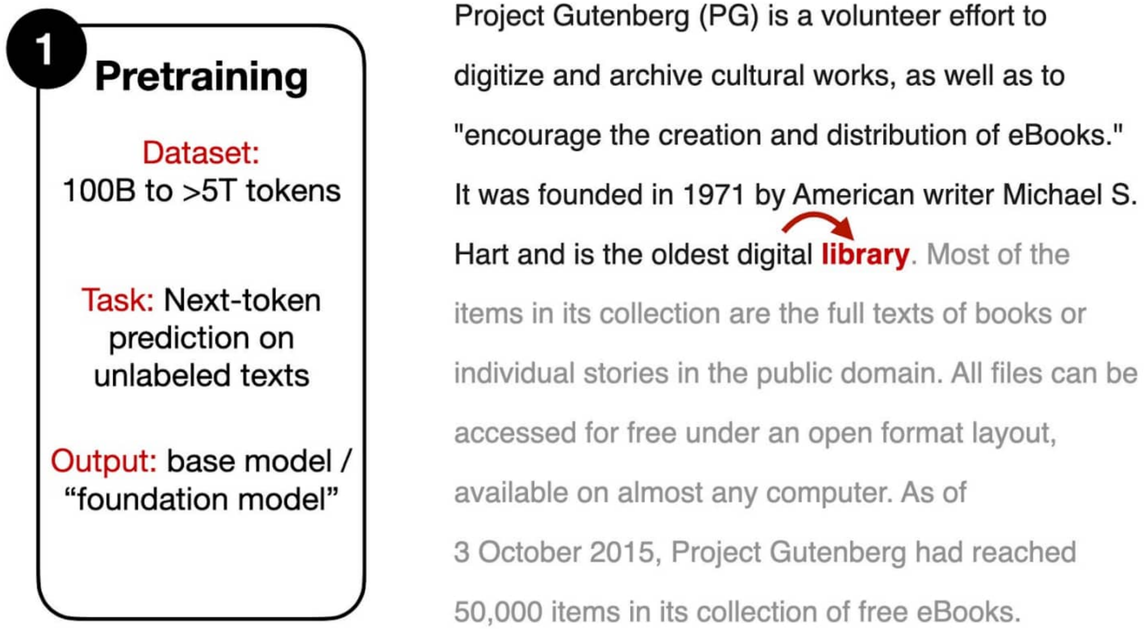
Self-Attention Module in GPT





Source: Jay Alammar

How to train LLMs?



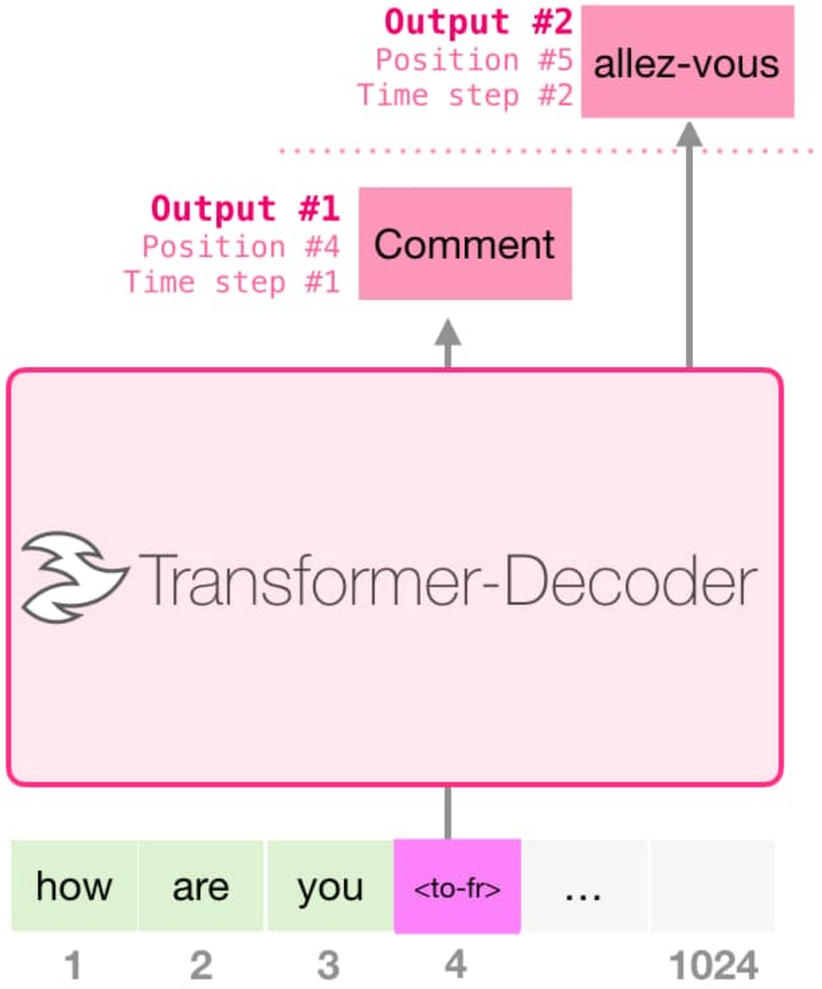
Project Gutenberg (PG) is a volunteer effort to digitize and archive cultural works, as well as to "encourage the creation and distribution of eBooks." It was founded in 1971 by American writer Michael S. Hart and is the oldest digital **library**. Most of the items in its collection are the full texts of books or individual stories in the public domain. All files can be accessed for free under an open format layout, available on almost any computer. As of 3 October 2015, Project Gutenberg had reached 50,000 items in its collection of free eBooks.

Source: Sebastian Raschka

How LLMs can solve tasks?

Training Dataset

I	am	a	student	<to-fr>	je	suis	étudiant
let	them	eat	cake	<to-fr>	Qu'ils	mangent	de
good	morning	<to-fr>	Bonjour				

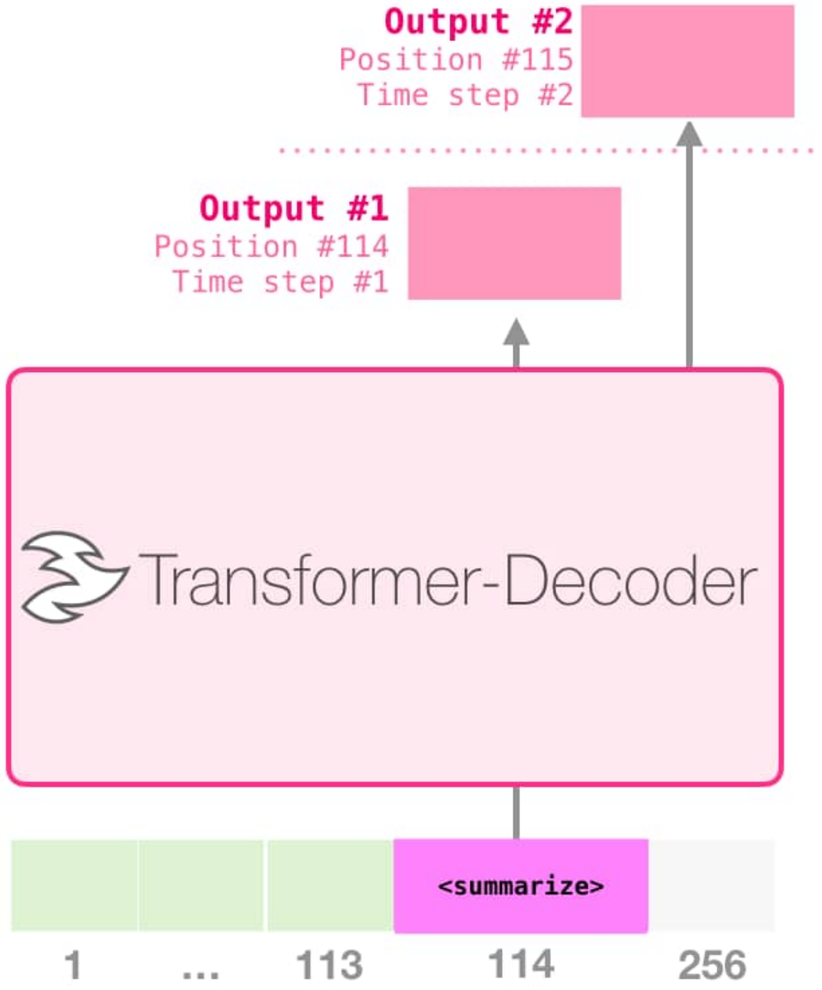


Source: Jay Alammar

How LLMs can solve tasks?

Training Dataset

Article #1 tokens		<summarize>	Article #1 Summary
Article #2 tokens	<summarize>	Article #2 Summary	padding
Article #3 tokens		<summarize>	Article #3 Summary



Source: Jay Alammar

How to train LLMs?

1 Pretraining

Dataset:
100B to >5T tokens

Task: Next-token prediction on unlabeled texts

Output: base model / "foundation model"

Project Gutenberg (PG) is a volunteer effort to digitize and archive cultural works, as well as to "encourage the creation and distribution of eBooks." It was founded in 1971 by American writer Michael S. Hart and is the oldest digital **library**. Most of the items in its collection are the full texts of books or individual stories in the public domain. All files can be accessed for free under an open format layout, available on almost any computer. As of 3 October 2015, Project Gutenberg had reached 50,000 items in its collection of free eBooks.

2 Supervised finetuning

More next-token prediction

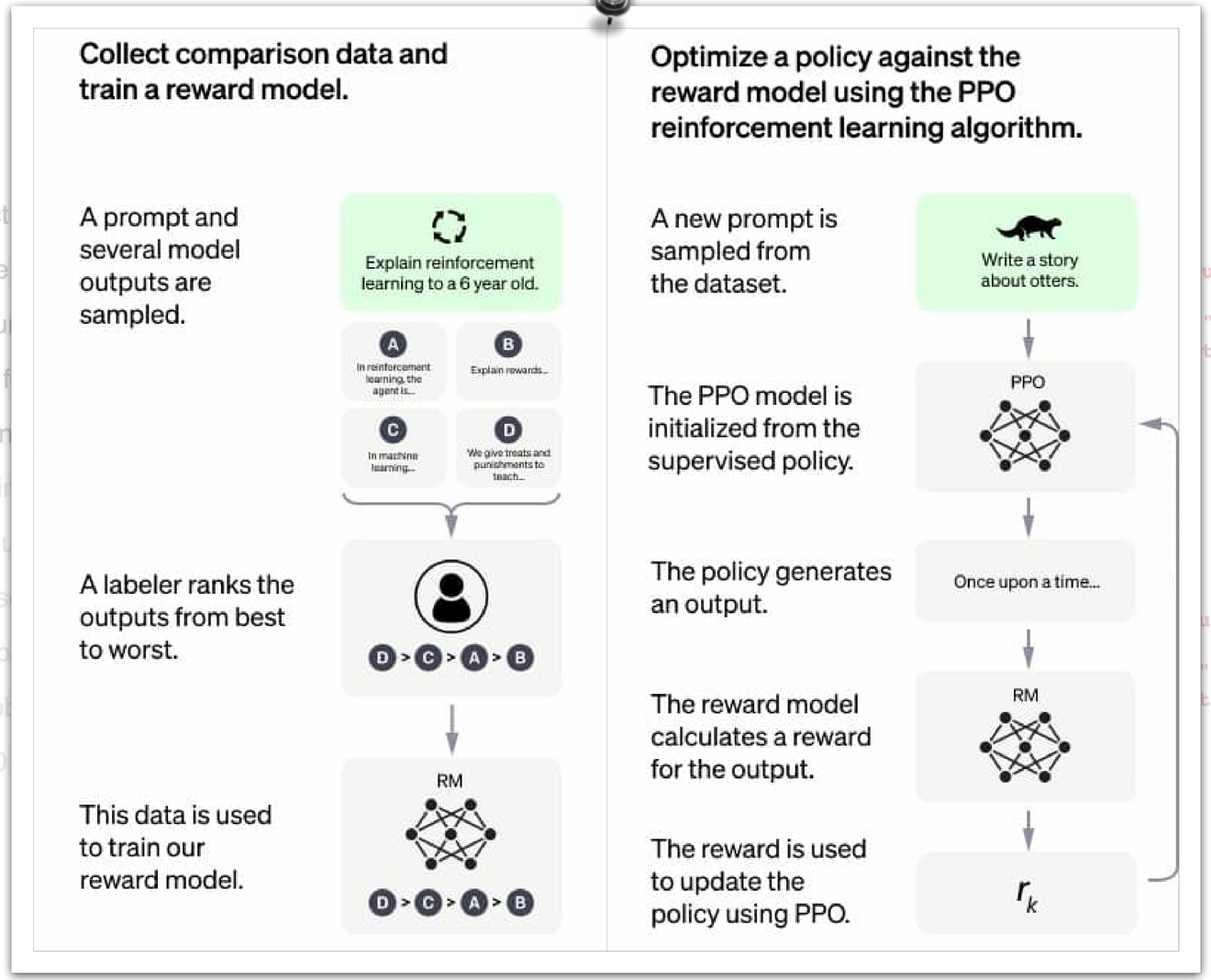
Usually 1k-50k instruction-response pairs

```
{
    "instruction": "Write a limerick about a
                    pelican.",
    "input": "",
    "output": "There once was a pelican so fine,
              \nHis beak was as colorful as
              sunshine,\nHe would fish all day,\nIn
              a very unique way,\nThis pelican was
              truly divine!\n\n\n"
},
```

```
{
    "instruction": "Identify the odd one out from
                    the group.",
    "input": "Carrot, Apple, Banana, Grape",
    "output": "Carrot\n\n"
},
```

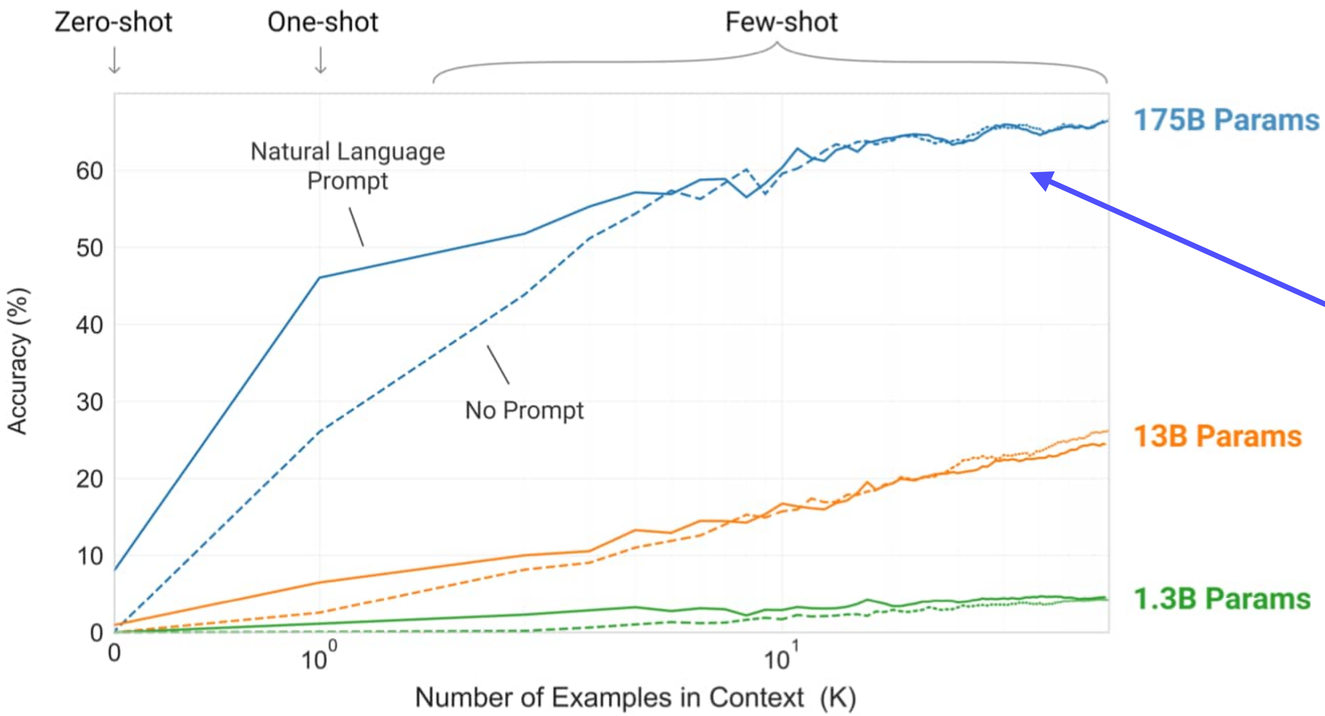
Source: Sebastian Raschka

How to train LLMs?

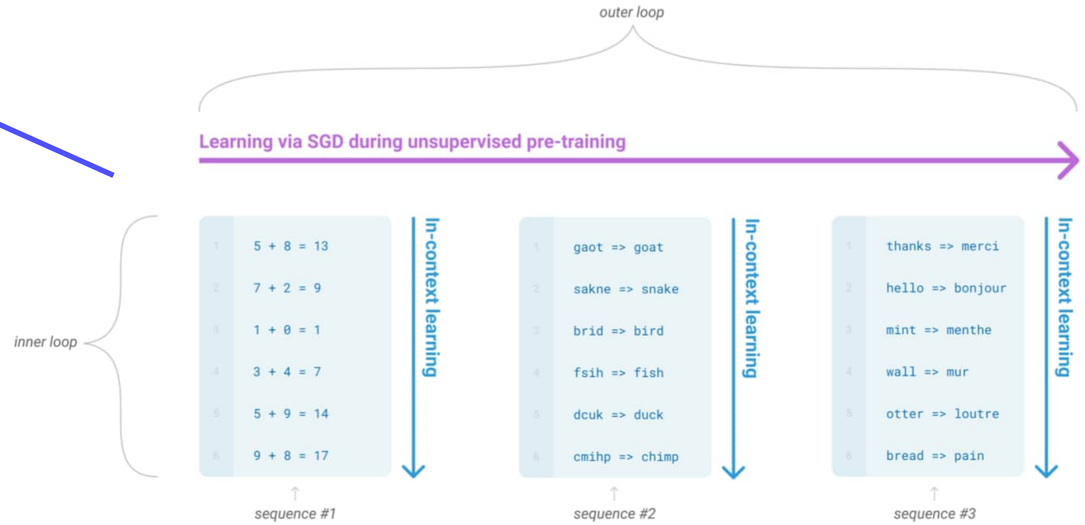


Source: Sebastian Raschka

LLMs are *in-context* Learners



require 355 years and
$4.5M to train (Tesla V100)
+ trained with 499 billion tokens



without fine-tuning, LLMs achieve improved performance with in-context demonstration

Brown et al., Language Models are Few-Shot Learners, ***NeurIPS*** **(2020)**

Prior-Fitted Networks

Sample *prior* datasets $\mathcal{D}^{(k)} \sim p(\mathcal{D})$

$$\mathcal{D}^{(1)} = \mathcal{D}^{(1)}_{\text{train}} \cup \{(x^{(1)}_{\text{test}}, y^{(1)}_{\text{test}})\}$$

$$\vdots$$

$$\mathcal{D}^{(K)} = \mathcal{D}^{(K)}_{\text{train}} \cup \{(x^{(K)}_{\text{test}}, y^{(K)}_{\text{test}})\}$$

Train the Prior-Fitted Net by minimizing

$$\text{prior-data NLL} - \sum_{k=1}^{K} \log q_\theta(y^{(k)}_{\text{test}} | x^{(k)}_{\text{test}}, \mathcal{D}^{(k)}_{\text{train}})$$

Actual training dataset and test input

$$(x_{\text{test}}, \mathcal{D}_{\text{train}})$$

Bayesian inference via the trained Prior-Fitted Net, with the actual training data and a test point as input

$$q_{\theta_*}(y_{\text{test}} | x_{\text{test}}, \mathcal{D}_{\text{train}}) \approx p(y_{\text{test}} | x_{\text{test}}, \mathcal{D}_{\text{train}})$$

Müller et al., Transformers Can Do Bayesian Inference, *ICLR* (2022)

Prior-Fitted Networks

$$p(y|x,\mathcal{D}) \propto \int_{\mathcal{T}} p(y|x,\tau)p(\mathcal{D}|\tau)p(\mathrm{d}\tau)$$

posterior predictive | likelihood | prior over task

approximate in KL-divergence

Sample *prior* datasets $\mathcal{D}^{(k)} \sim p(\mathcal{D})$

$$\mathcal{D}^{(1)} = \mathcal{D}^{(1)}_{\text{train}} \cup \{(x^{(1)}_{\text{test}}, y^{(1)}_{\text{test}})\}$$

$$\vdots$$

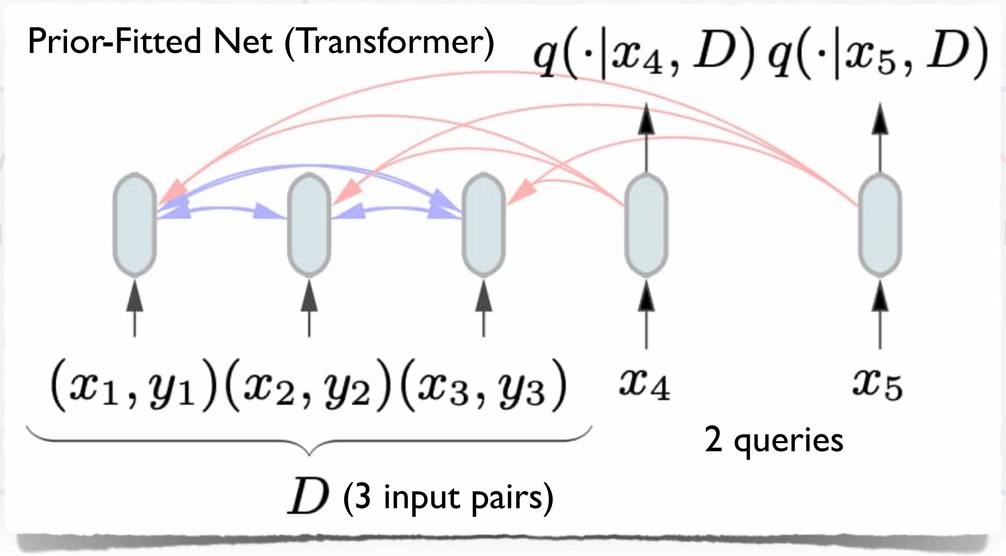
$$\mathcal{D}^{(K)} = \mathcal{D}^{(K)}_{\text{train}} \cup \{(x^{(K)}_{\text{test}}, y^{(K)}_{\text{test}})\}$$

Train the Prior-Fitted Net by minimizing

$$\text{prior-data NLL} - \sum_{k=1}^{K} \log q_\theta(y^{(k)}_{\text{test}} | x^{(k)}_{\text{test}}, \mathcal{D}^{(k)}_{\text{train}})$$

Actual training dataset and test input

$$(x_{\text{test}}, \mathcal{D}_{\text{train}})$$



Müller et al., Transformers Can Do Bayesian Inference, *ICLR* (2022)

Prior-Fitted Networks

$$p(y|x,\mathcal{D}) \propto \int_{\mathcal{T}} p(y|x,\tau)p(\mathcal{D}|\tau)p(\mathrm{d}\tau)$$

posterior predictive | likelihood | prior over task

approximate in KL-divergence

Sample *prior* datasets $\mathcal{D}^{(k)} \sim p(\mathcal{D})$

$$\mathcal{D}^{(1)} = \mathcal{D}^{(1)}_{\text{train}} \cup \{(x^{(1)}_{\text{test}}, y^{(1)}_{\text{test}})\}$$

$$\vdots$$

$$\mathcal{D}^{(K)} = \mathcal{D}^{(K)}_{\text{train}} \cup \{(x^{(K)}_{\text{test}}, y^{(K)}_{\text{test}})\}$$

Train the Prior-Fitted Net by minimizing

$$\text{prior-data NLL} - \sum_{k=1}^{K} \log q_\theta(y^{(k)}_{\text{test}}|x^{(k)}_{\text{test}}, \mathcal{D}^{(k)}_{\text{train}})$$

Prior-Fitted Net with parameter θ_*

Actual training dataset and test input

$(x_{\text{test}}, \mathcal{D}_{\text{train}})$

Predict y_{test}

Bayesian inference via the trained Prior-Fitted Net, with the actual training data and a test point as input

$$q_{\theta_*}(y_{\text{test}}|x_{\text{test}}, \mathcal{D}_{\text{train}}) \approx p(y_{\text{test}}|x_{\text{test}}, \mathcal{D}_{\text{train}})$$

Müller et al., Transformers Can Do Bayesian Inference, *ICLR* (2022)

Transformer for Tabular Data

Published as a conference paper at ICLR 2023

TABPFN: A TRANSFORMER THAT SOLVES SMALL TABULAR CLASSIFICATION PROBLEMS IN A SECOND

Noah Hollmann[*,1,2] **Samuel Müller**[*,1] **Katharina Eggensperger**[1] **Frank Hutter**[1,3]
[1] University of Freiburg, [2] Charité University Medicine Berlin
[3] Bosch Center for Artificial Intelligence * Equal contribution.
Correspondence to noah.hollmann@charite.de & muellesa@cs.uni-freiburg.de

인공지능/딥러닝 최고권위학회 ***ICLR*** (2023)
Spotlight (top 25%) 선정 발표

nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › articles › article

Article | Open access | Published: 08 January 2025

Accurate predictions on small data with a tabular foundation model

Noah Hollmann ✉, Samuel Müller ✉, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister & Frank Hutter ✉

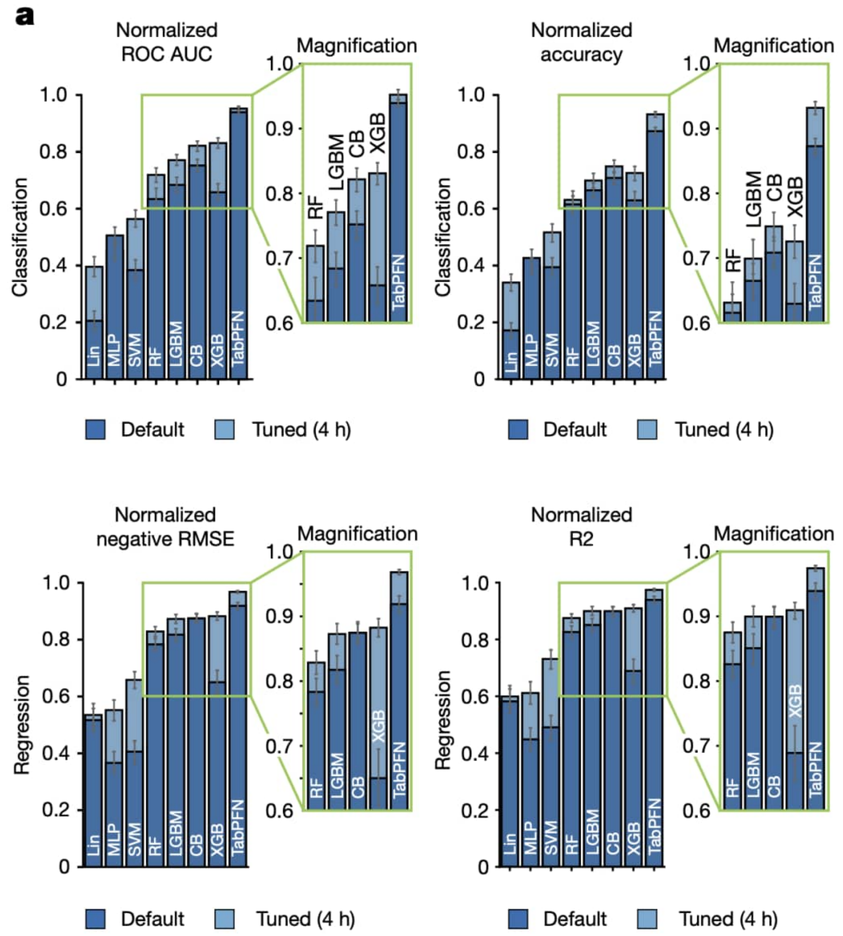
Nature **637**, 319–326 (2025) | Cite this article

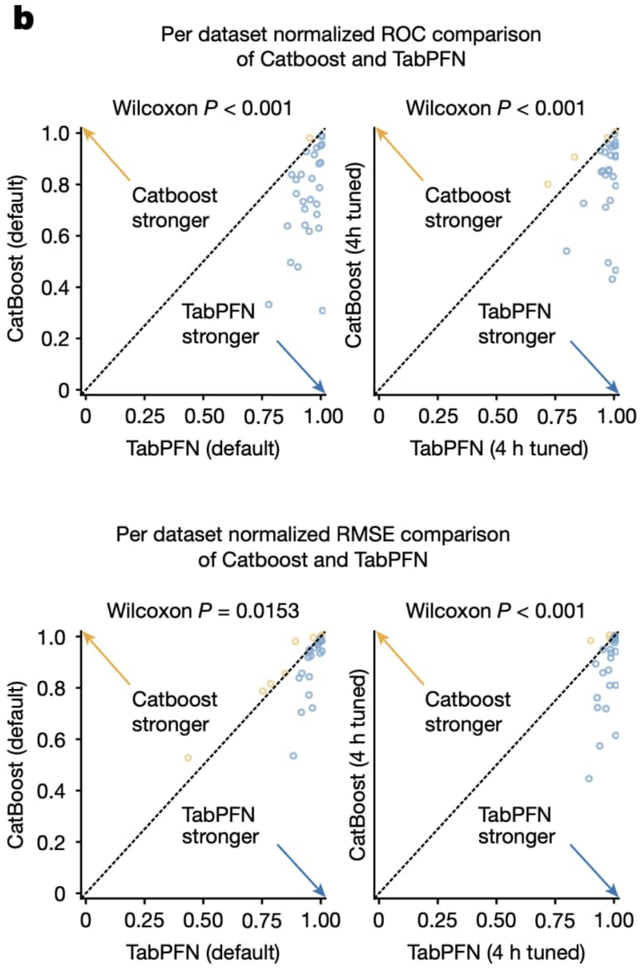
223k Accesses | **381** Altmetric | Metrics

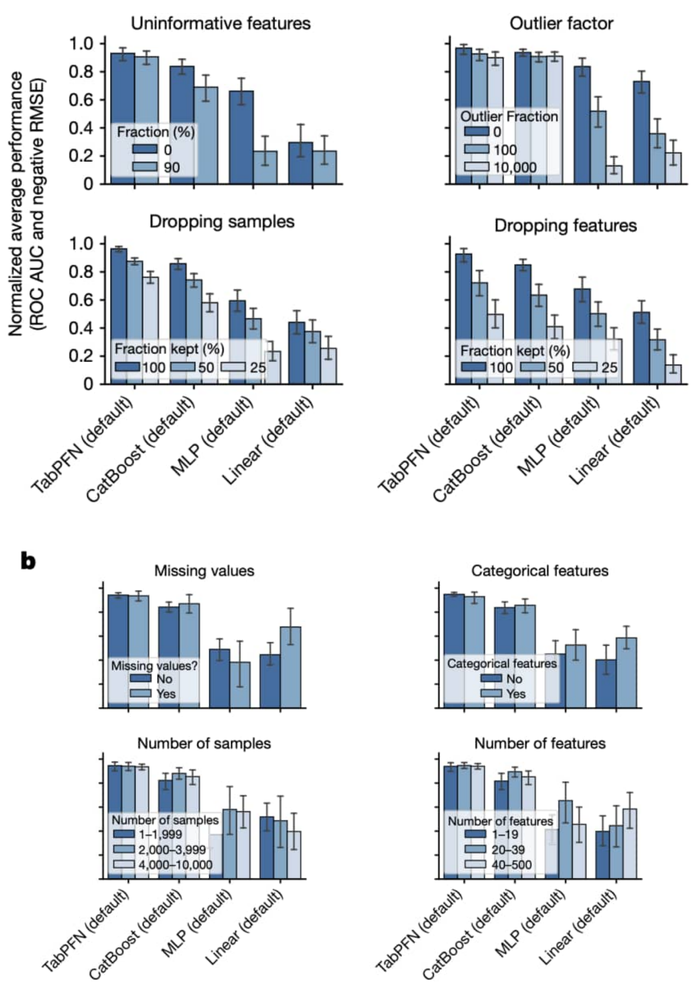
올해 1월 Nature (2025) 본지에 게재

Hollmann et al., Accurate predictions on small data with a tabular foundation model, ***Nature*** (2025)

TabPFN is more accurate and robust





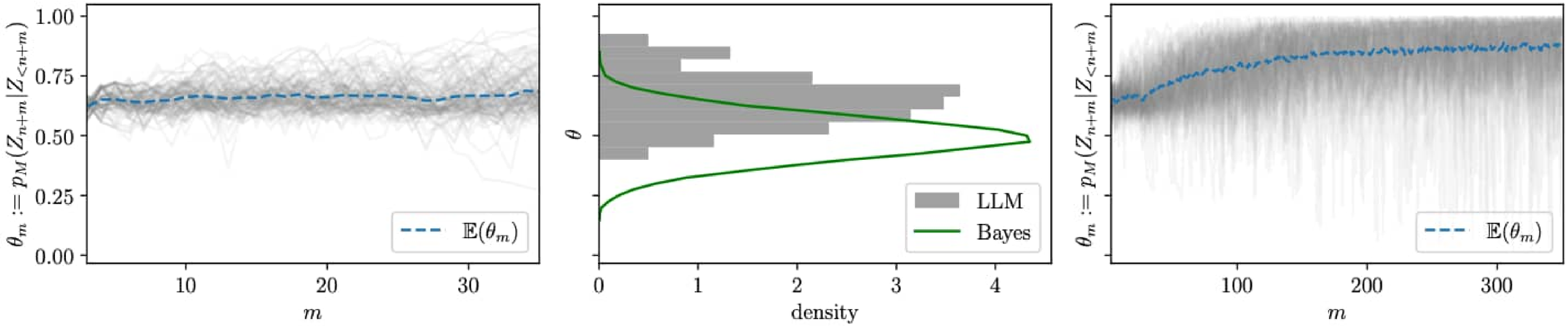


Hollmann et al., Accurate predictions on small data with a tabular foundation model, *Nature* (2025)

ICL in LLMs is ***not*** Bayesian

Some people believe LLMs can do this
but many AI scientists disagree

- Many researchers have postulated ICL as approximately Bayesian inference
- Recent study presents an empirical evidence that LLMs violate the martingale property which is a necessary condition for exchangeability → ***not*** Bayesian!
  - uncertainty of an LLM's predictive distribution remains opaque

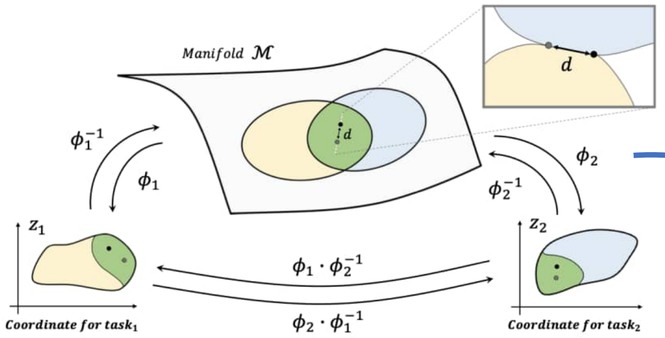


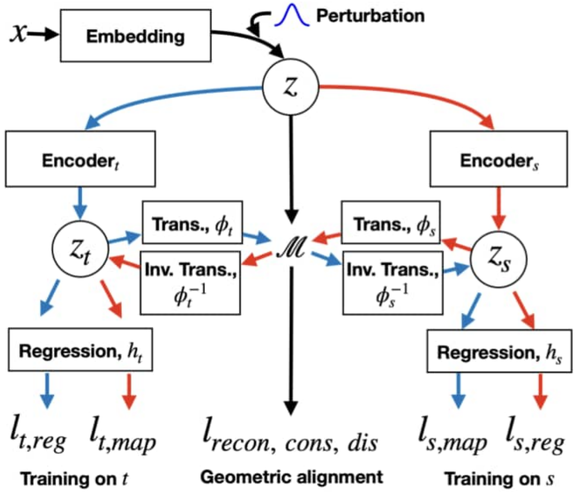
Falck et al., Is In-Context Learning in Large Language Models Bayesian? A Martingale Perspective, ***ICML*** **(2024)**

Part 4

Foundation Model in AI4Science

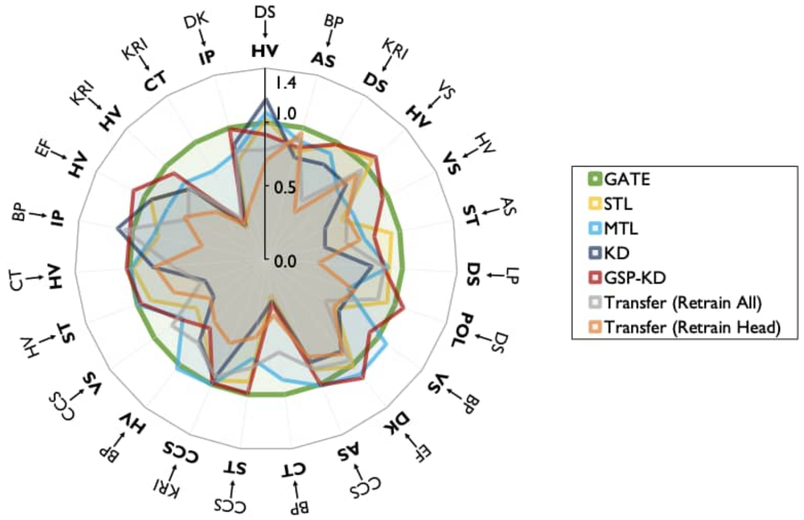
Toward Scalable Multi-task Learning in AI4Science





GATE (Ko et al., ICLR 2024)

Representation Learning via GATE

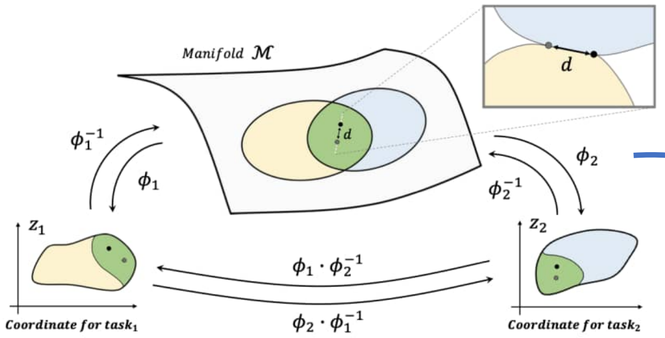


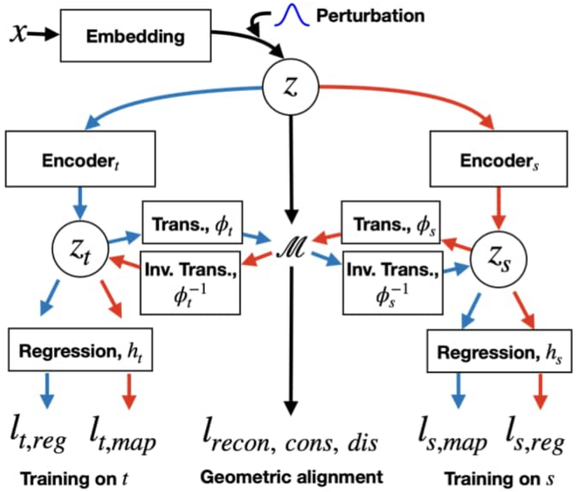
Multi-task learning in chemical property prediction *without* domain knowledge has limitation in scalability

Ko et al., Geometrically Aligned Transfer Encoder for Inductive Transfer in Regression Tasks, ***ICLR*** **(2024)**

Lee et al., Scalable Multi-Task Transfer Learning for Molecular Property Prediction, ***ICML AI4Science Workshop*** **(2024)**

Toward Scalable Multi-task Learning in AI4Science

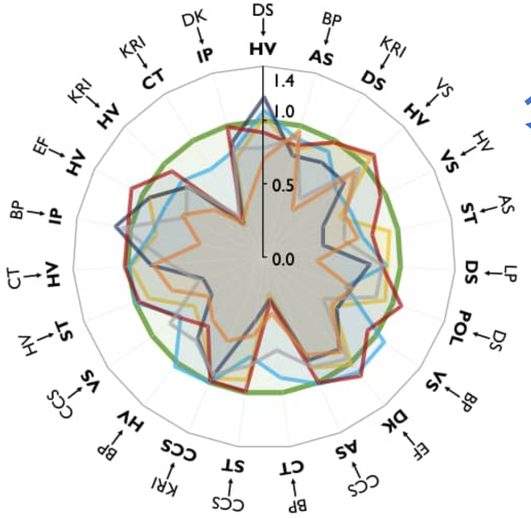




GATE (Ko et al., ICLR 2024)

Representation Learning via GATE

Meta learning can solve many adaptation problems in this field



40 tasks!

Algorithm 1 Bi-level Optimization for GATE

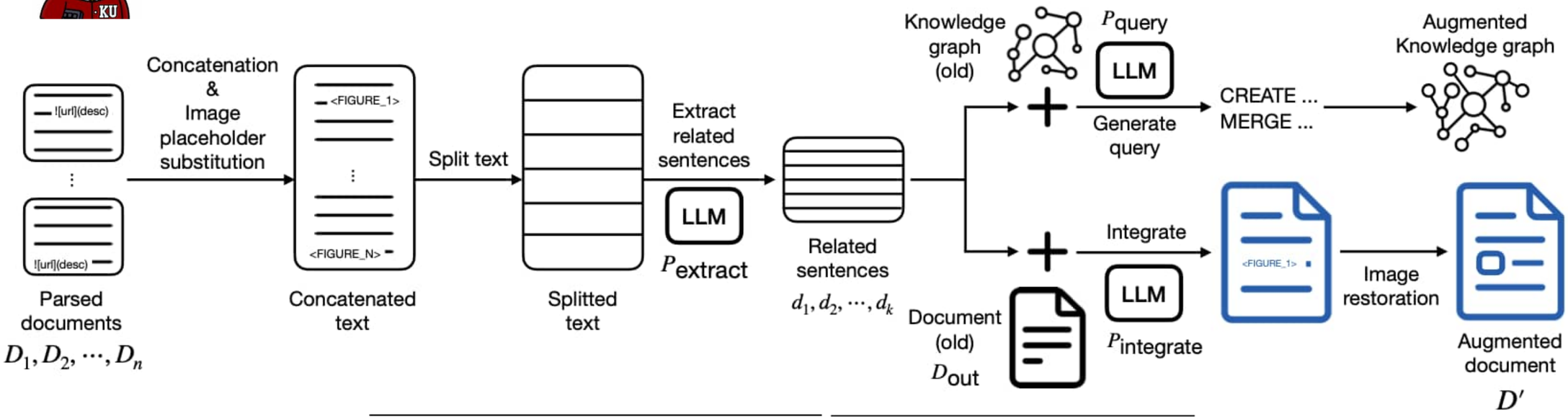
Input: Training data D_{tr}, validation data D_{val}
Initialize model θ, transfer ratio λ, transfer momentum m, v
repeat
  $\theta \leftarrow \arg\min_{\theta} L(D_{tr}, \theta, \lambda)$; Inner loop
  $\lambda \leftarrow \arg\min_{\lambda} L(D_{val}, \theta, \lambda, m, v)$; Outer loop
until converged

Data-driven transfer learning via bilevel optim. (Lee et al., 2024)

Ko et al., Geometrically Aligned Transfer Encoder for Inductive Transfer in Regression Tasks, *ICLR* (2024)
Lee et al., Scalable Multi-Task Transfer Learning for Molecular Property Prediction, *ICML AI4Science Workshop* (2024)

Toward Scalable Multi-task Learning in AI4Science

Data engineering is a key to enhance the performance of models!

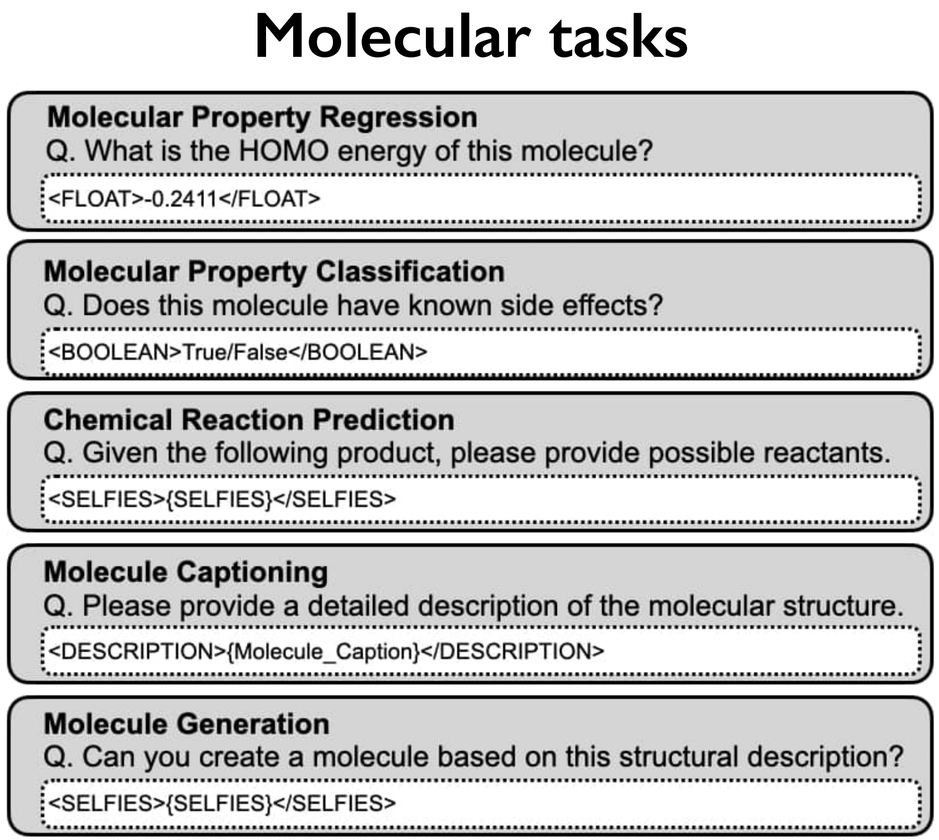


	Gain↑	Gain per token (K)↑
Wikidata+Pandoc	254/158	1.382
Wikidata+html2markdown	**337**/143	0.690
Wikidata+Markdownify	308/122	2.050
Wikidata+html2text	269/116	1.981
Wikidata+BS	246/146	3.276
Wikidata+ReactionParser	293/**166**	**6.810**

	Gain↑	Gain per token (M)↑
LLM only	0.000	-
LLM+Pandoc	0.078	2.345
LLM+html2markdown	**0.099**	1.344
LLM+html2text	0.062	2.581
LLM+Markdownify	0.087	3.885
LLM+BS	0.059	3.942
LLM+ReactionParser	0.075	**9.235**

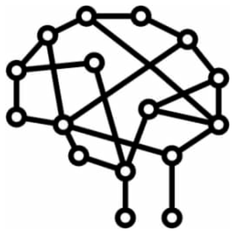
Ko et al., Filling in the Gaps: LLM-based structured data generation from semi-structured scientific data, *ICML AI4Science workshop*

Foundation Model for Molecular Tasks



Foundation Model for Molecular Tasks

how can I answer these questions?



Answer

Foundation Model

Molecular tasks

Molecular Property Regression
Q. What is the HOMO energy of this molecule?
<FLOAT>-0.2411</FLOAT>

Molecular Property Classification
Q. Does this molecule have known side effects?
<BOOLEAN>True/False</BOOLEAN>

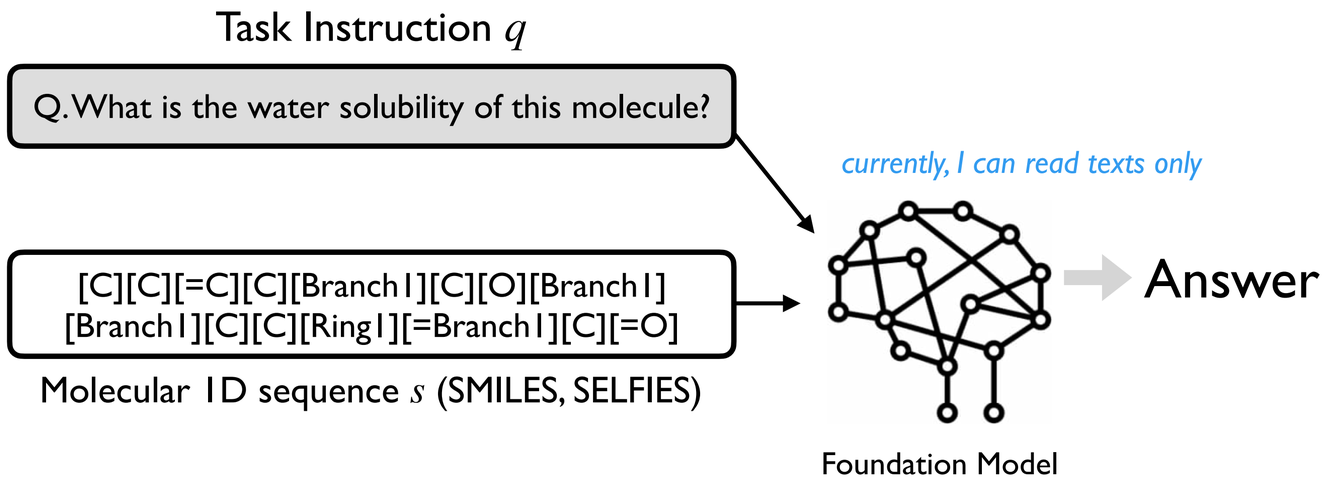
Chemical Reaction Prediction
Q. Given the following product, please provide possible reactants.
<SELFIES>{SELFIES}</SELFIES>

Molecule Captioning
Q. Please provide a detailed description of the molecular structure.
<DESCRIPTION>{Molecule_Caption}</DESCRIPTION>

Molecule Generation
Q. Can you create a molecule based on this structural description?
<SELFIES>{SELFIES}</SELFIES>

Foundation Model for Molecular Tasks

Molecular Instruction Tuning



Molecular tasks

Molecular Property Regression
Q. What is the HOMO energy of this molecule?
<FLOAT>-0.2411</FLOAT>

Molecular Property Classification
Q. Does this molecule have known side effects?
<BOOLEAN>True/False</BOOLEAN>

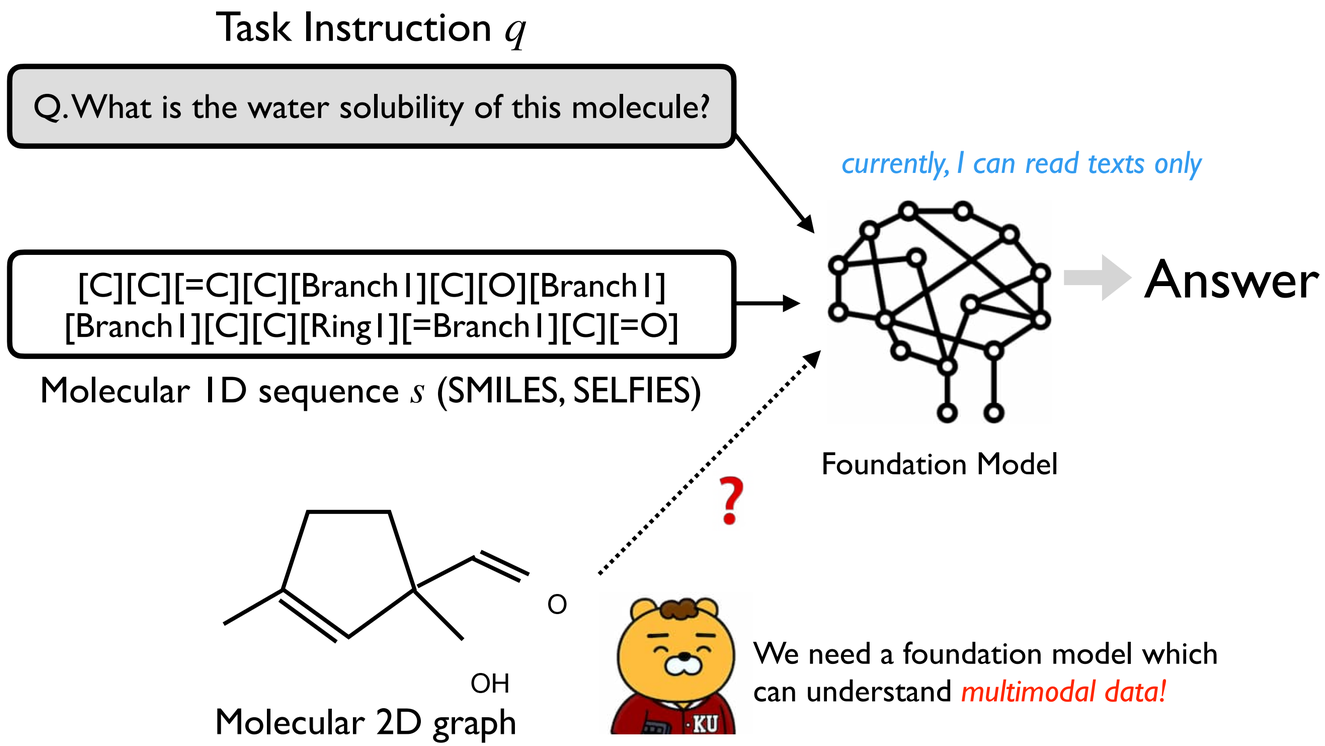
Chemical Reaction Prediction
Q. Given the following product, please provide possible reactants.
<SELFIES>{SELFIES}</SELFIES>

Molecule Captioning
Q. Please provide a detailed description of the molecular structure.
<DESCRIPTION>{Molecule_Caption}</DESCRIPTION>

Molecule Generation
Q. Can you create a molecule based on this structural description?
<SELFIES>{SELFIES}</SELFIES>

Foundation Model for Molecular Tasks

Molecular Instruction Tuning



Molecular tasks

Molecular Property Regression
Q. What is the HOMO energy of this molecule?
<FLOAT>-0.2411</FLOAT>

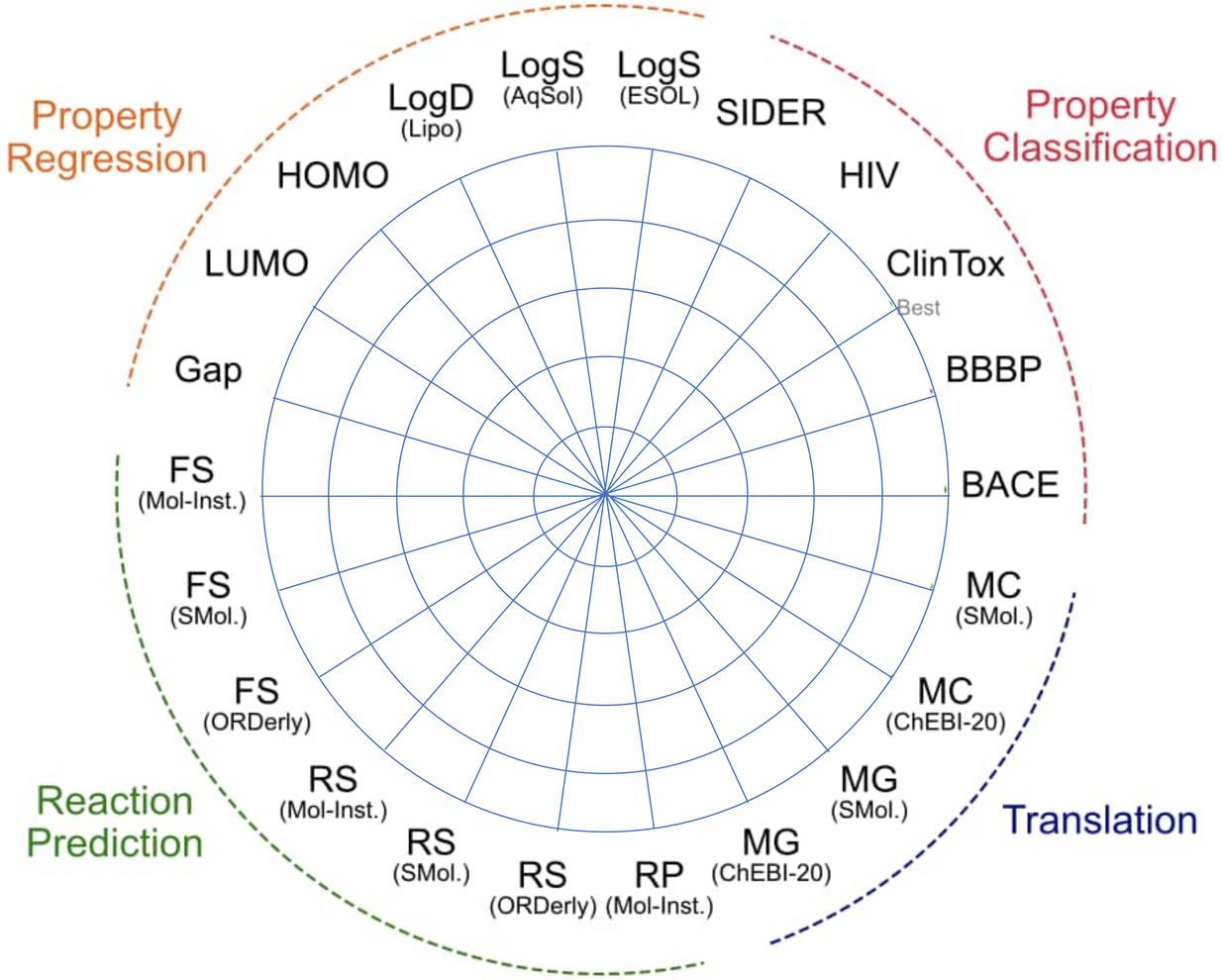
Molecular Property Classification
Q. Does this molecule have known side effects?
<BOOLEAN>True/False</BOOLEAN>

Chemical Reaction Prediction
Q. Given the following product, please provide possible reactants.
<SELFIES>{SELFIES}</SELFIES>

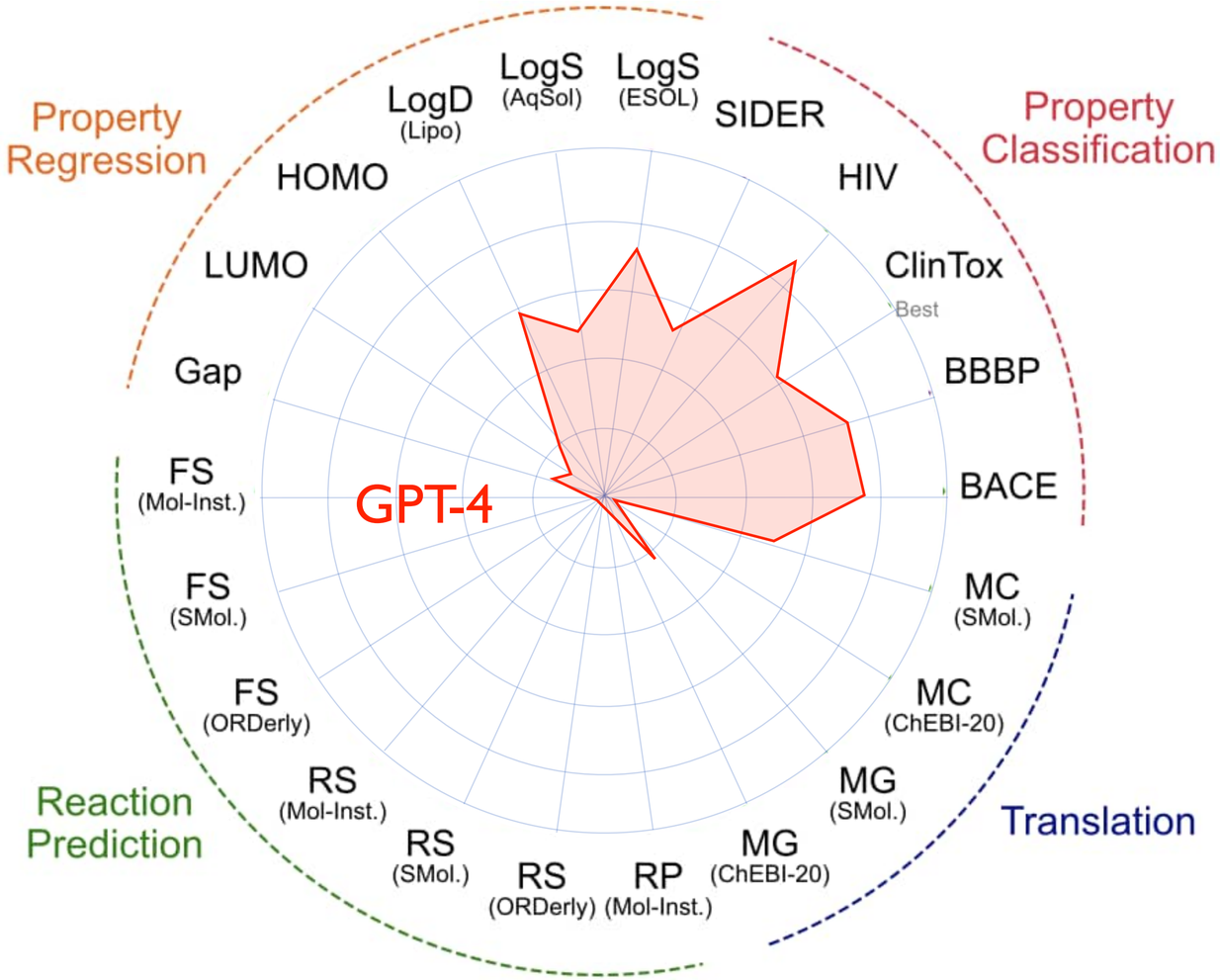
Molecule Captioning
Q. Please provide a detailed description of the molecular structure.
<DESCRIPTION>{Molecule_Caption}</DESCRIPTION>

Molecule Generation
Q. Can you create a molecule based on this structural description?
<SELFIES>{SELFIES}</SELFIES>

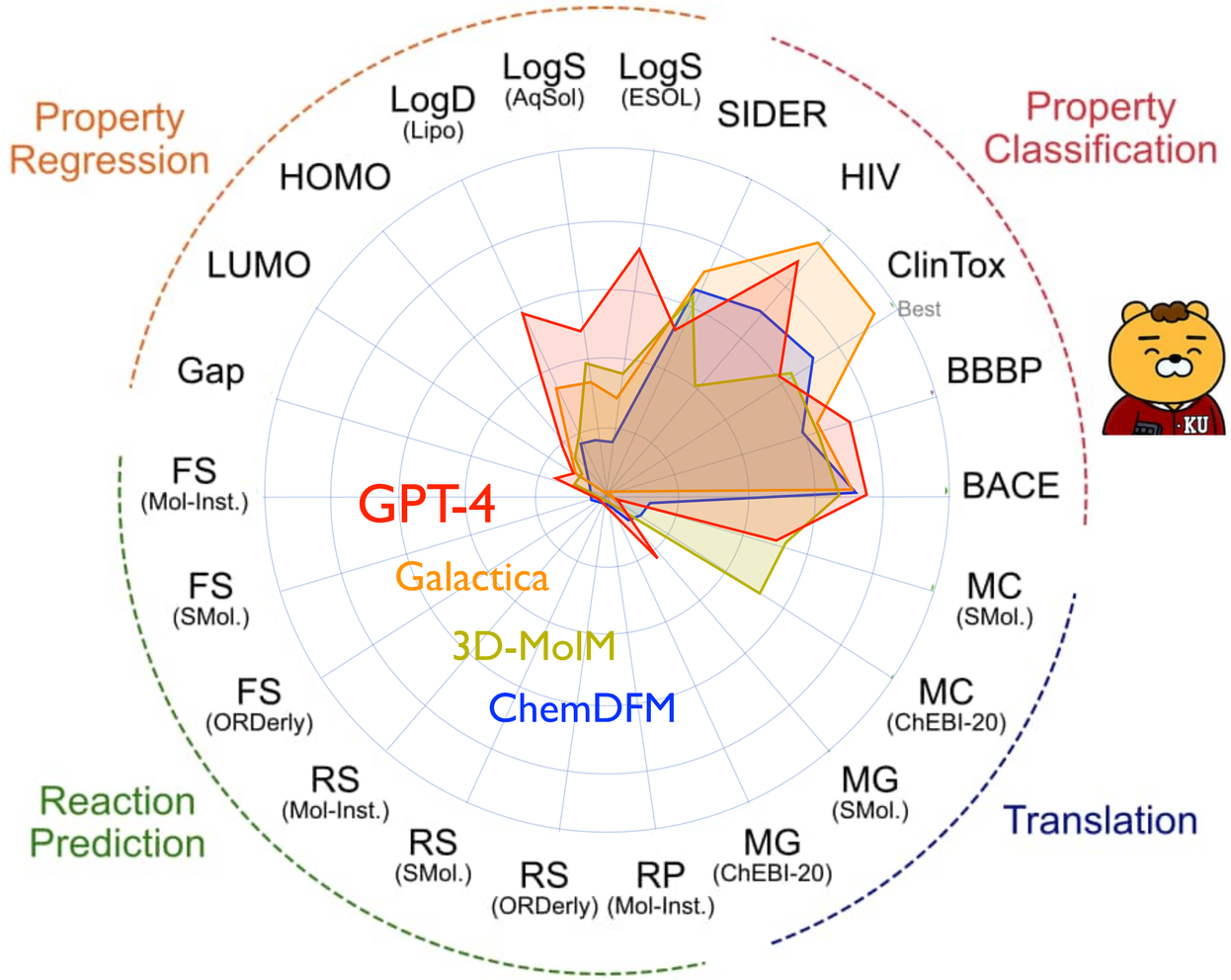
State of Generalist Models



State of Generalist Models

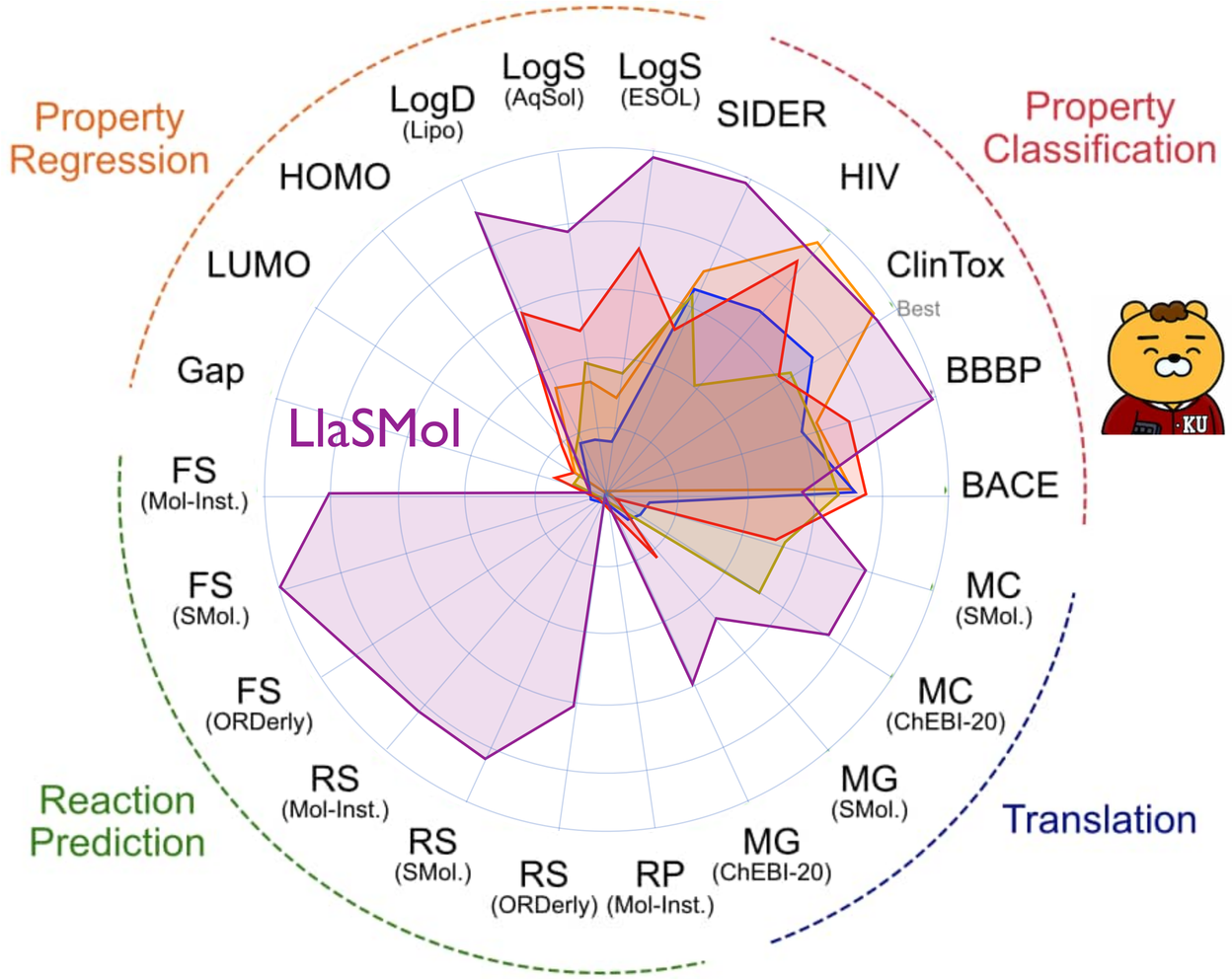


State of Generalist Models



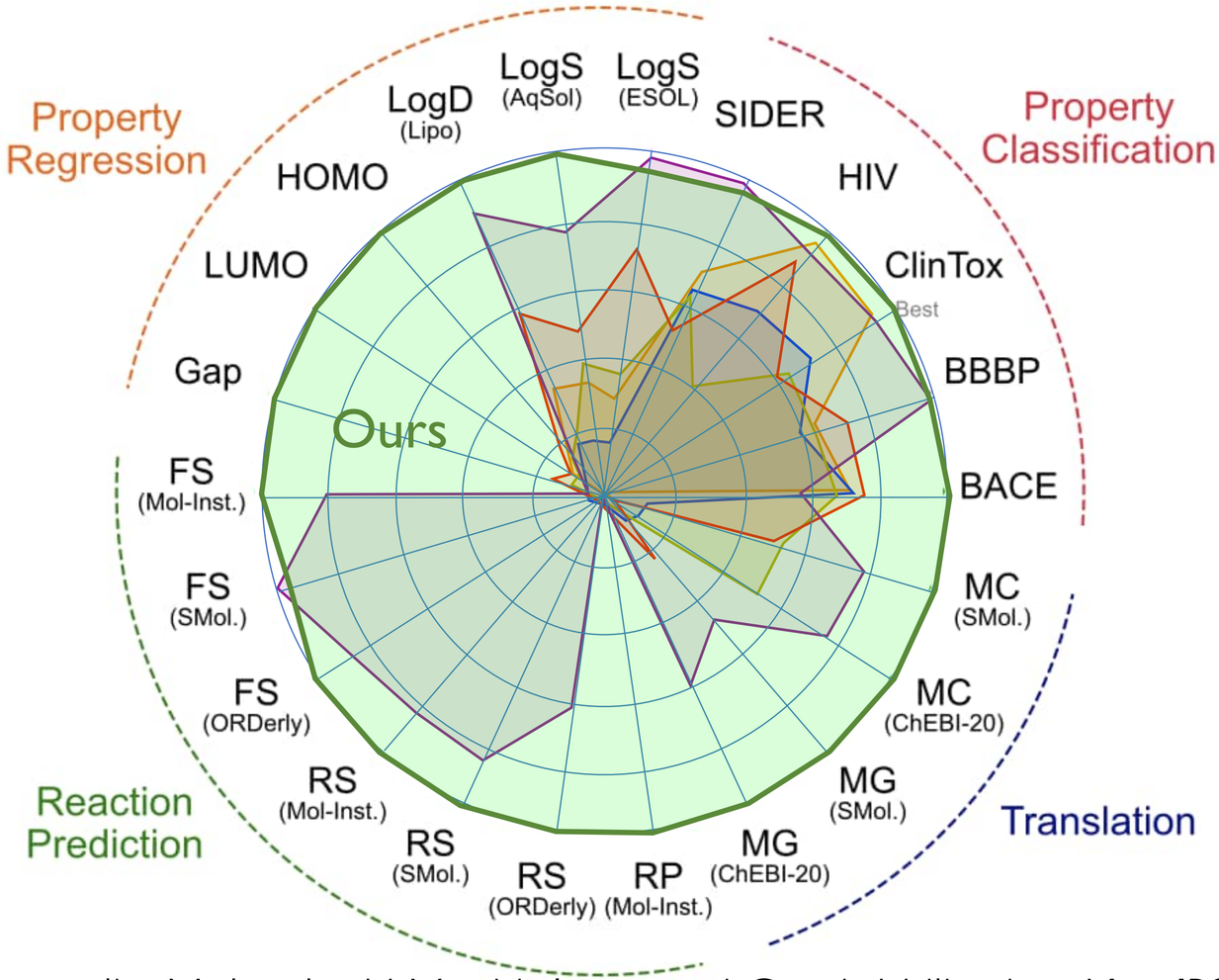
Note that 3D-MolM is a *multimodal* LLM, but it is inferior to seq-based generalist models

State of Generalist Models



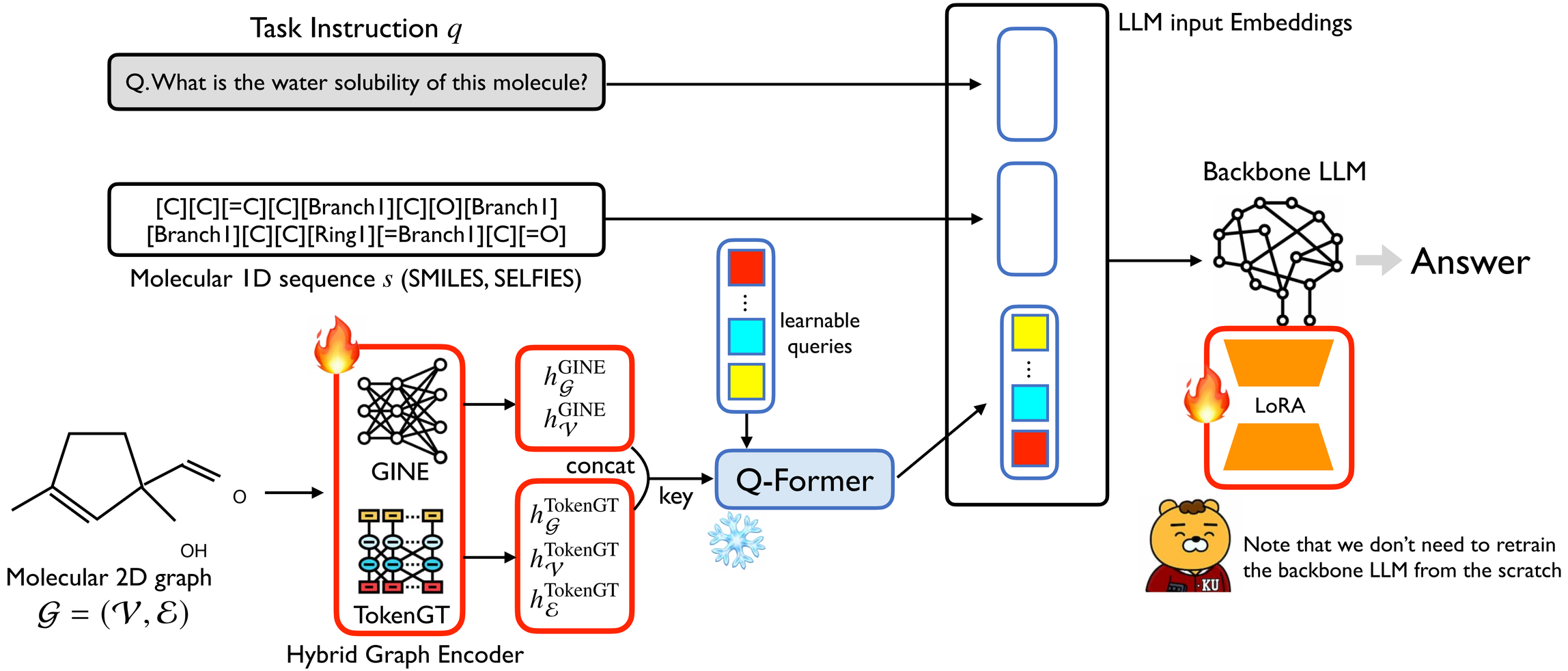
LlaSMol shows good performance
but it is still a seq-based unimodal LLM

State of Generalist Models & Ours (Mol-LLM)



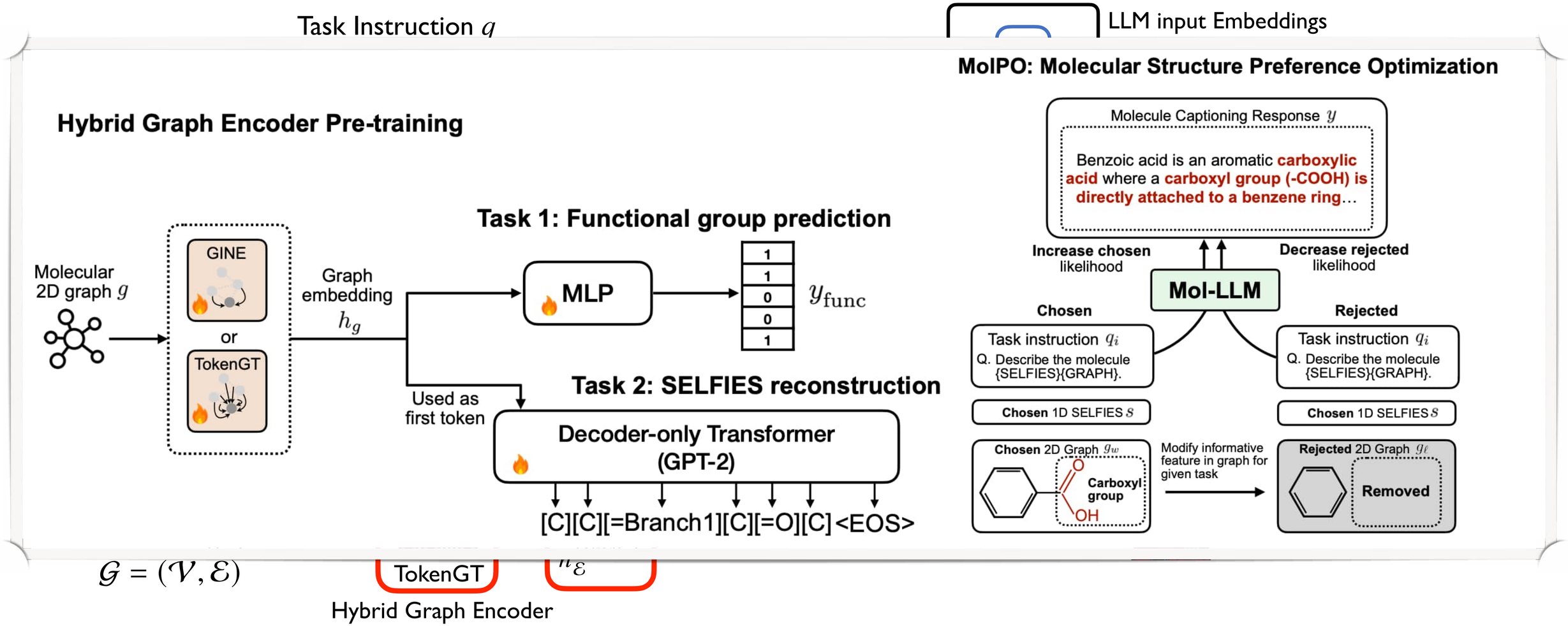
Lee et al., Mol-LLM: Multimodal Generalist Molecular LLM with Improved Graph Utilization, *NeurIPS AI4Science Workshop* (2025)

Multimodal Molecular LLM on Graph & Text



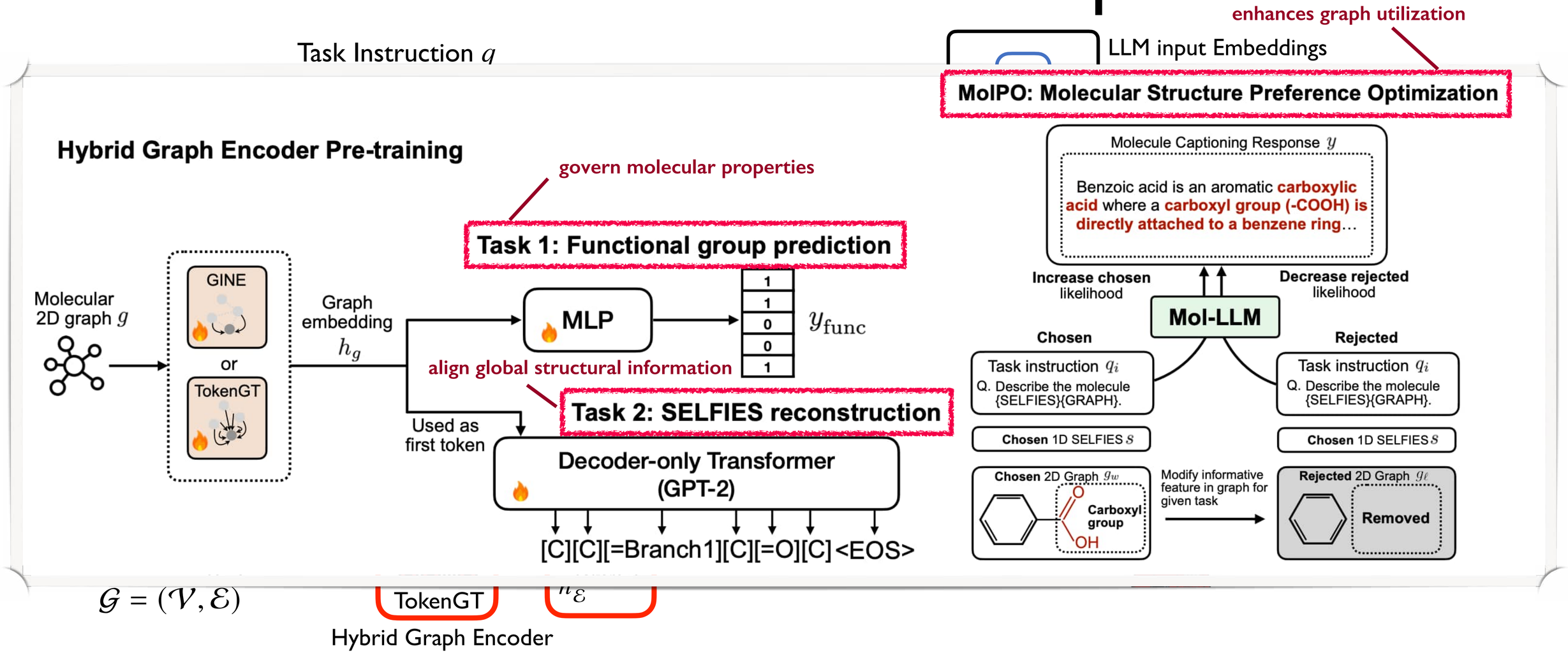
Lee et al., Mol-LLM: Multimodal Generalist Molecular LLM with Improved Graph Utilization, *NeurIPS AI4Science Workshop* (2025)

Multimodal Molecular LLM on Graph & Text

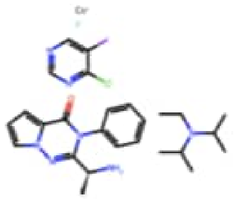
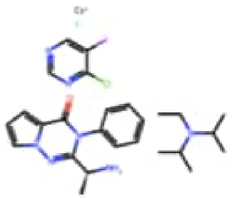
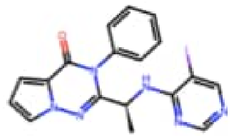
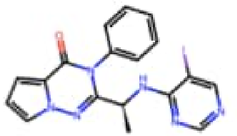
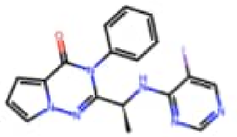
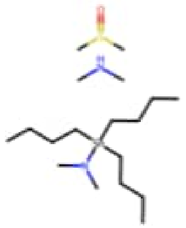
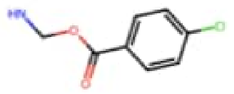
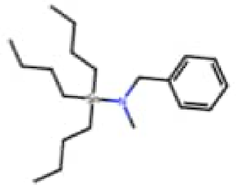
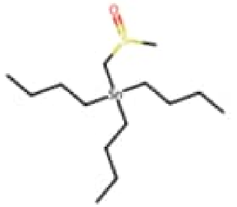
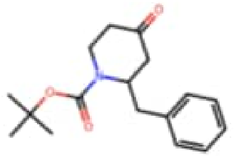
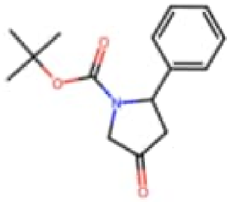
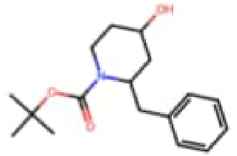
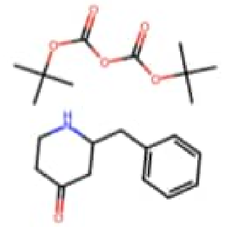
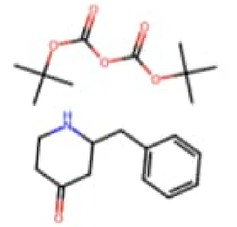
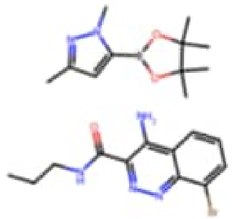
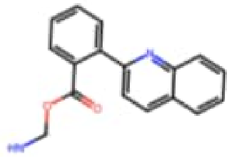
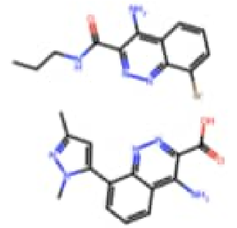
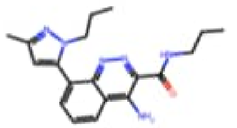
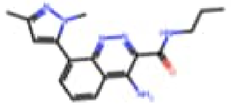


Lee et al., Mol-LLM: Multimodal Generalist Molecular LLM with Improved Graph Utilization, *NeurIPS AI4Science Workshop* (2025)

Multimodal Molecular LLM on Graph & Text



Lee et al., Mol-LLM: Multimodal Generalist Molecular LLM with Improved Graph Utilization, *NeurIPS AI4Science Workshop* (2025)

	Question	INPUT_MOLECULE	Galactica	LlaSMol	Mol-LLM (Ours)	Ground Truth
FS-InD	Please provide a feasible product that could be formed using these reactants and reagents: [INPUT_MOLECULE]					
FS-OOD	Please provide a feasible product that could be formed using these reactants and reagents: [INPUT_MOLECULE]					
RS-InD	Can you list the reactants that might result in the chemical product [INPUT_MOLECULE] ?					
RS-OOD	Can you list the reactants that might result in the chemical product [INPUT_MOLECULE] ?					

Lee et al., Mol-LLM: Multimodal Generalist Molecular LLM with Improved Graph Utilization, *NeurIPS AI4Science Workshop* (2025)

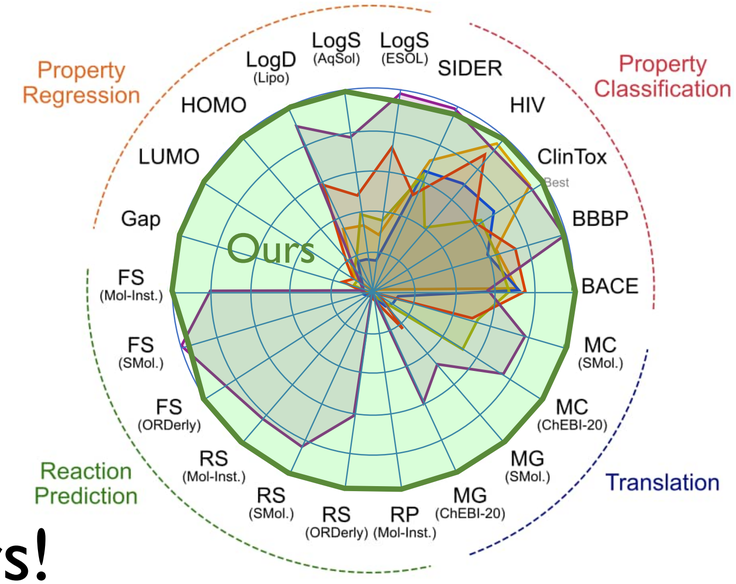
Final Part

Discussion & Open Problems

Conclusion & Discussion

- *Foundation Models* for *AI4Science* are promising for solving many tasks
  - property classification/regression, reaction prediction, generation etc.
- *Multimodal* approach is necessary for *generalist* models
  - test benchmark should be evaluated rigorously
- *Statistical* approach to foundation models is necessary
  - uncertainty quantification is a key for search method
- What is *Next*?

✍️ *Key message*: *software & data engineering* really matters!



Q&A

KOREA UNIVERSITY