CS181 Artificial Intelligence

Advanced Topics in AI

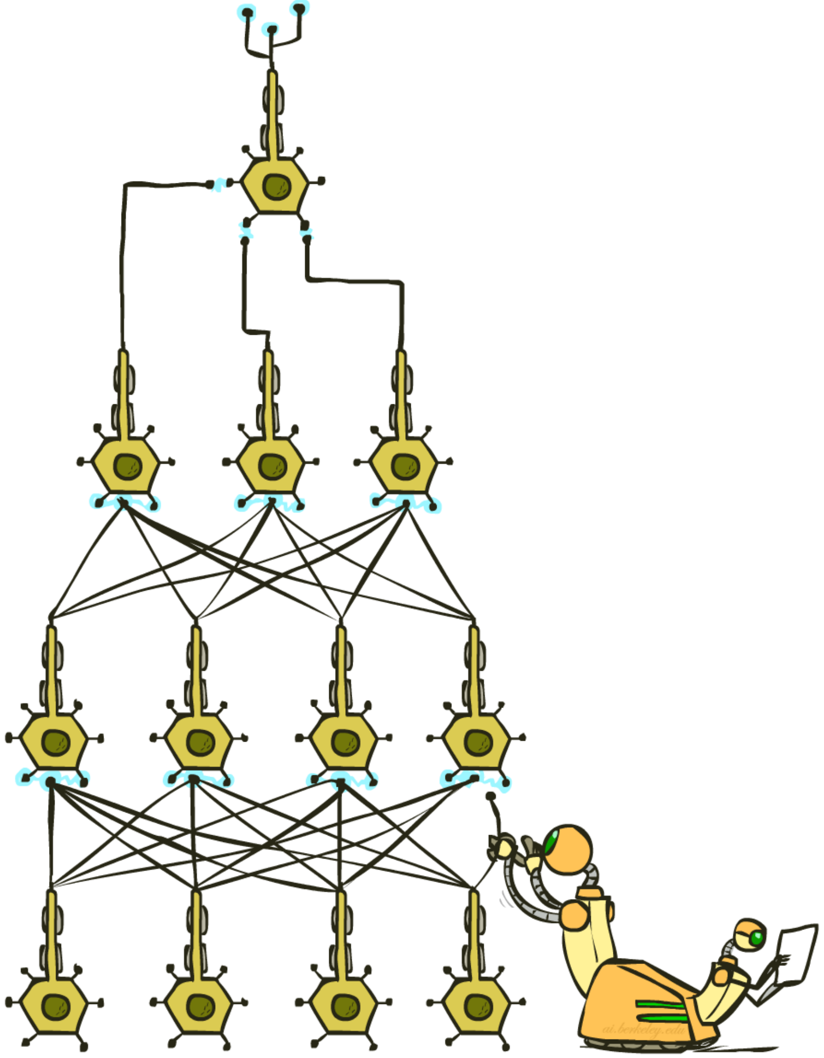
Kan Ren

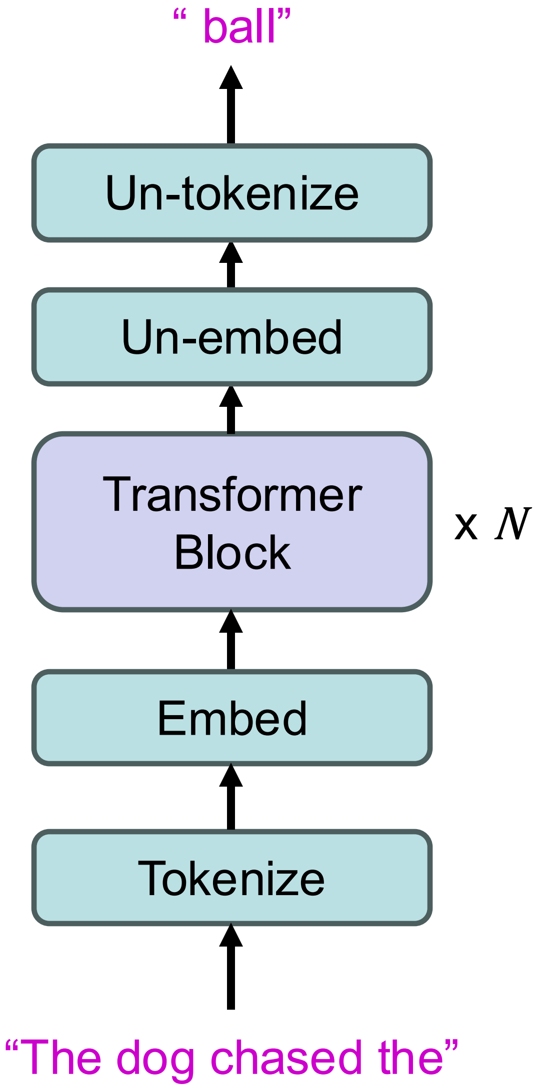
Fall 2024

Advanced topics

- Advanced deep learning
  - Transformer
  - Generative modeling
  - Foundation model
- Advanced reinforcement learning
  - Real-world reinforcement learning
  - AI agents
- Responsible AI

Recall of Transformer architecture

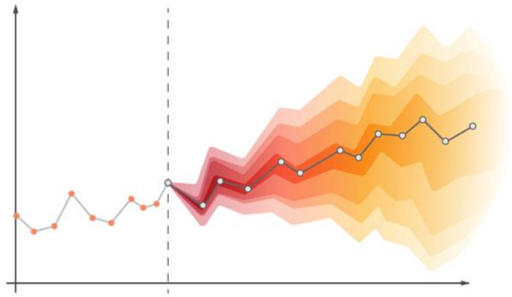




Is Transformer only good at language?

Transformer in different data modalities

- Transformer in 1D data

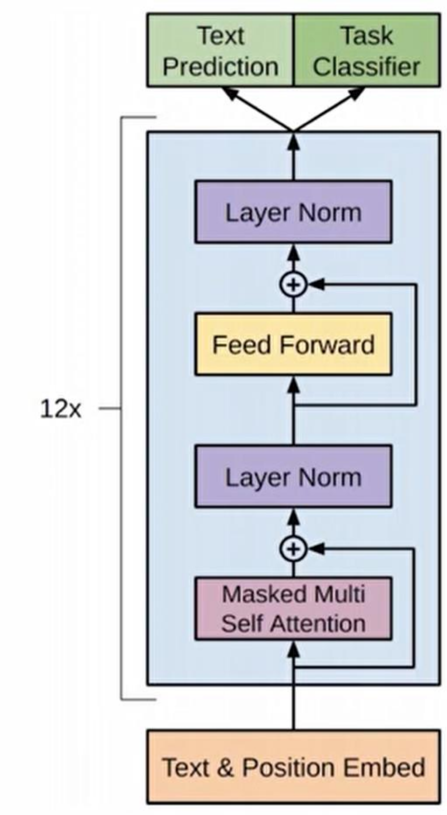


- Transformer in 2D data

- Transformer in 3D data

Transformer in 1D data

- GPT-1 (Radford et al 2018)

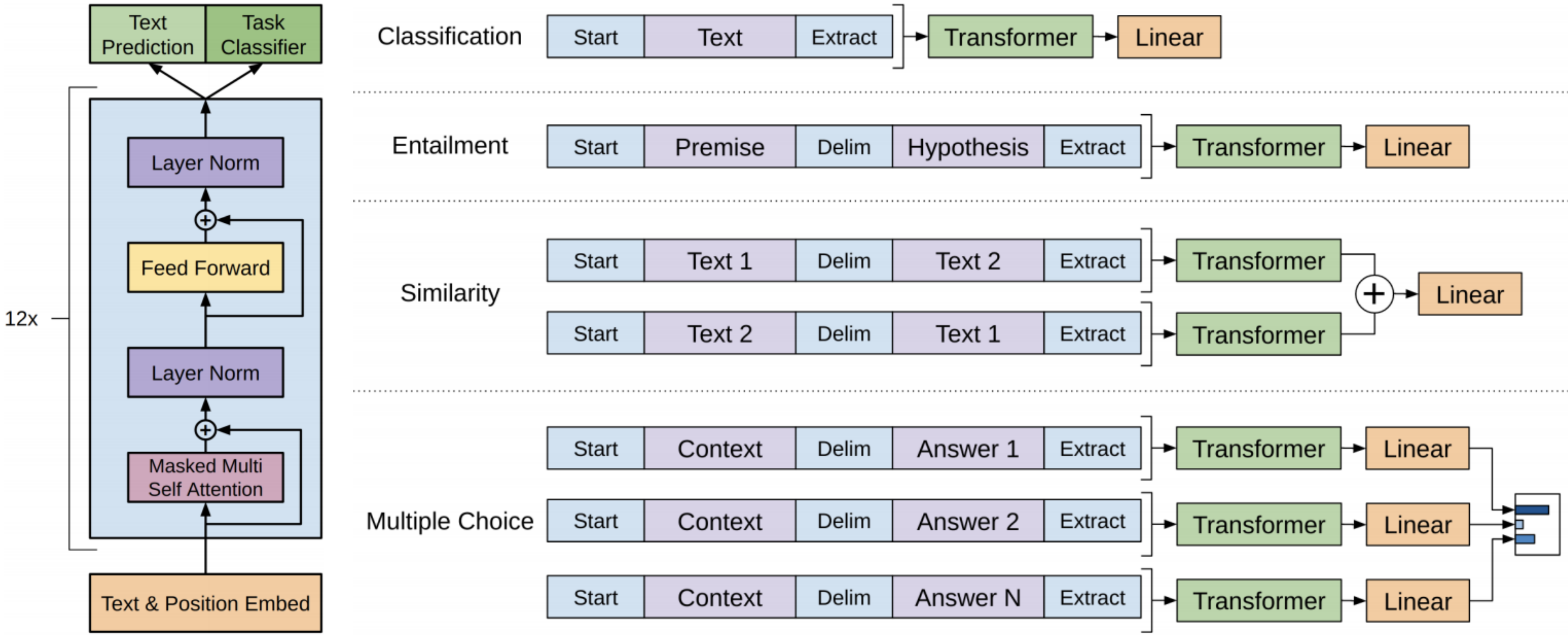


DATASET	TASK	SOTA	OURS
SNLI	Textual Entailment	89.3	**89.9**
MNLI Matched	Textual Entailment	80.6	**82.1**
MNLI Mismatched	Textual Entailment	80.1	**81.4**
SciTail	Textual Entailment	83.3	**88.3**
QNLI	Textual Entailment	82.3	**88.1**
RTE	Textual Entailment	**61.7**	56.0
STS-B	Semantic Similarity	81.0	**82.0**
QQP	Semantic Similarity	66.1	**70.3**
MRPC	Semantic Similarity	**86.0**	82.3
RACE	Reading Comprehension	53.3	**59.0**
ROCStories	Commonsense Reasoning	77.6	**86.5**
COPA	Commonsense Reasoning	71.2	**78.6**
SST-2	Sentiment Analysis	**93.2**	91.3
CoLA	Linguistic Acceptability	35.0	**45.4**
GLUE	Multi Task Benchmark	68.9	**72.8**

Radford, Alec. "Improving language understanding by generative pre-training." (2018).

Transformer in 1D data

- GPT-1 (Radford et al 2018)



Radford, Alec. "Improving language understanding by generative pre-training." (2018).

Transformer in 1D data

- GPT-2
- All the supervised learning are the subset of the unsupervised pretrain when the language model is large enough.

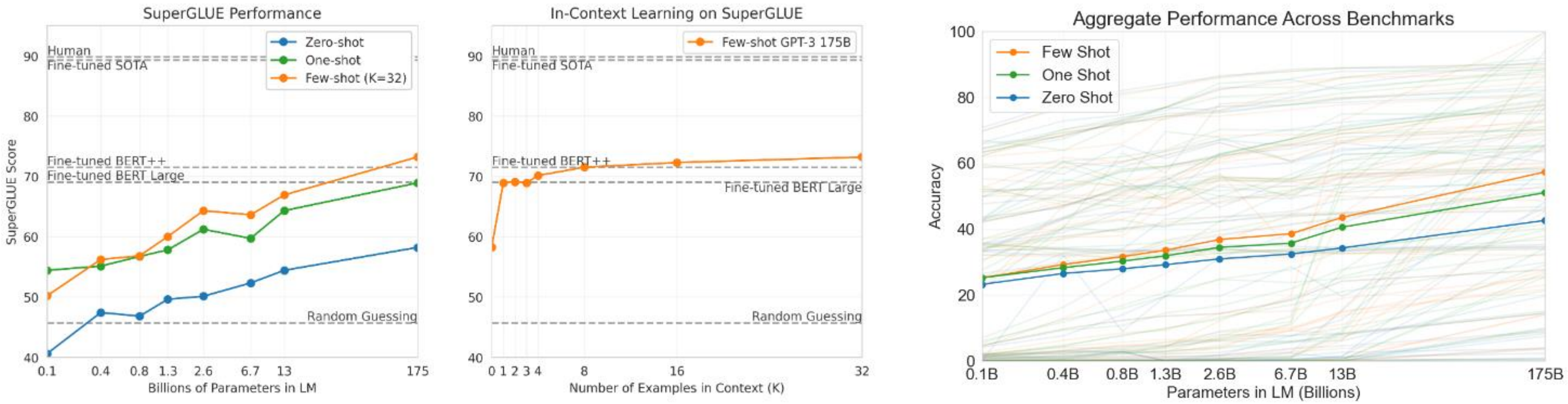
- Unsupervised task: $p(x) = \prod_{i=1}^{n} p(s_n | s_1, ..., s_{n-1}) \longrightarrow p(s_{n-k}, ..., s_n | s_1, ..., s_{n-k-1})$
- Supervised task: $p(output|input, task)$

<Task description> <input> <output>

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. "Language models are unsupervised multitask learners." *OpenAI blog* 1, no. 8 (2019): 9.

Transformer in 1D data

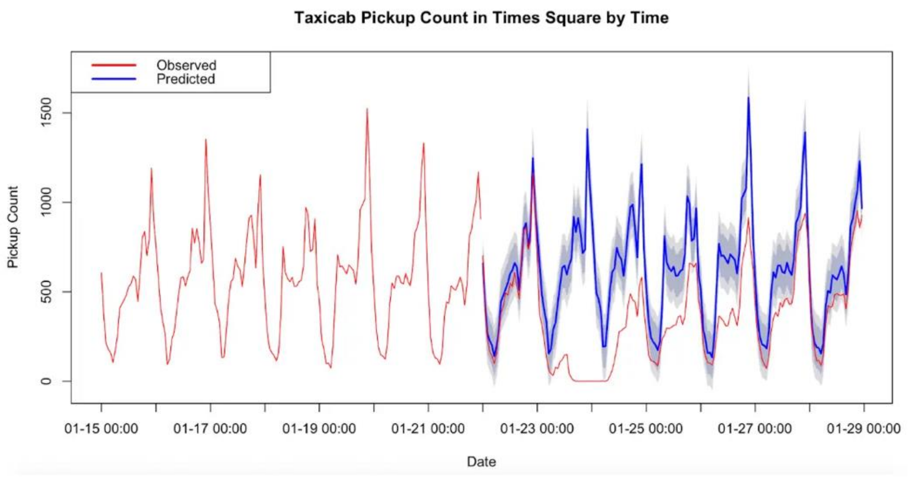
- GPT-3
- Model architecture: (sparse attention, more parameters)
  - 1.5B --> 175B
  - Dense attention, each token calculate the attention from all the other tokens
  - Sparse attention, each token calculate with the token that distance <k, 2k, 3k ......
  - Make sense that local information is dense
  - Reduce the complexity that make the LLM able to handle longer inputs
- Inference:
  - Zero-shot --> Few-shot
  - No need to finetune but few examples instead

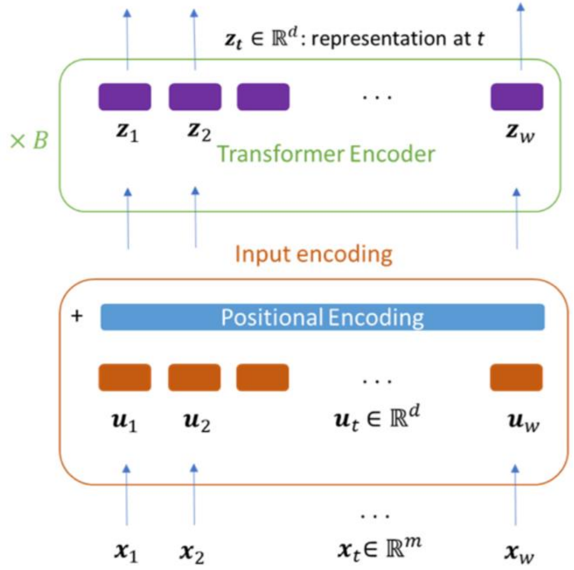


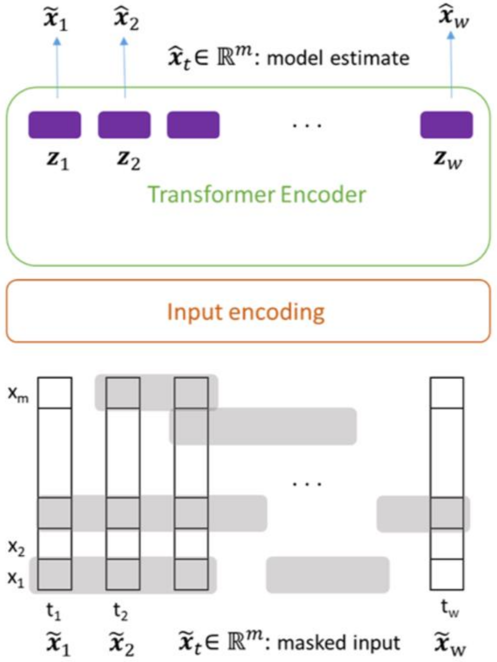
Brown, Tom et al. "Language models are few-shot learners." Advances in neural information processing systems 33 (2020): 1877-1901.

Transformer in 1D data

- Time-series Transformer (TST) with masked auto-encoder pretraining.



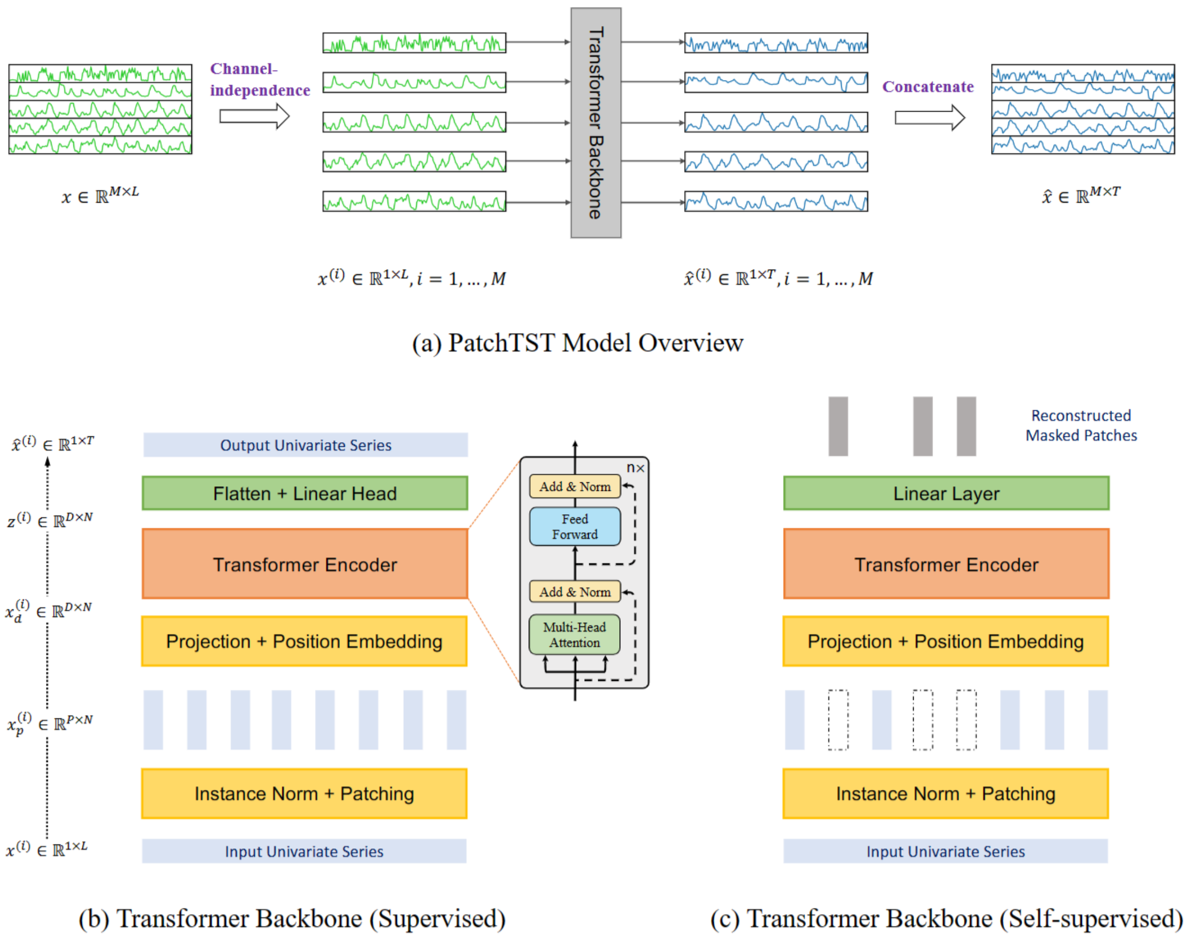




Zerveas et al, "A Transformer-based Framework for Multivariate Time Series Representation Learning", KDD 2021

Transformer in 1D data

- PatchTST



(a) PatchTST Model Overview

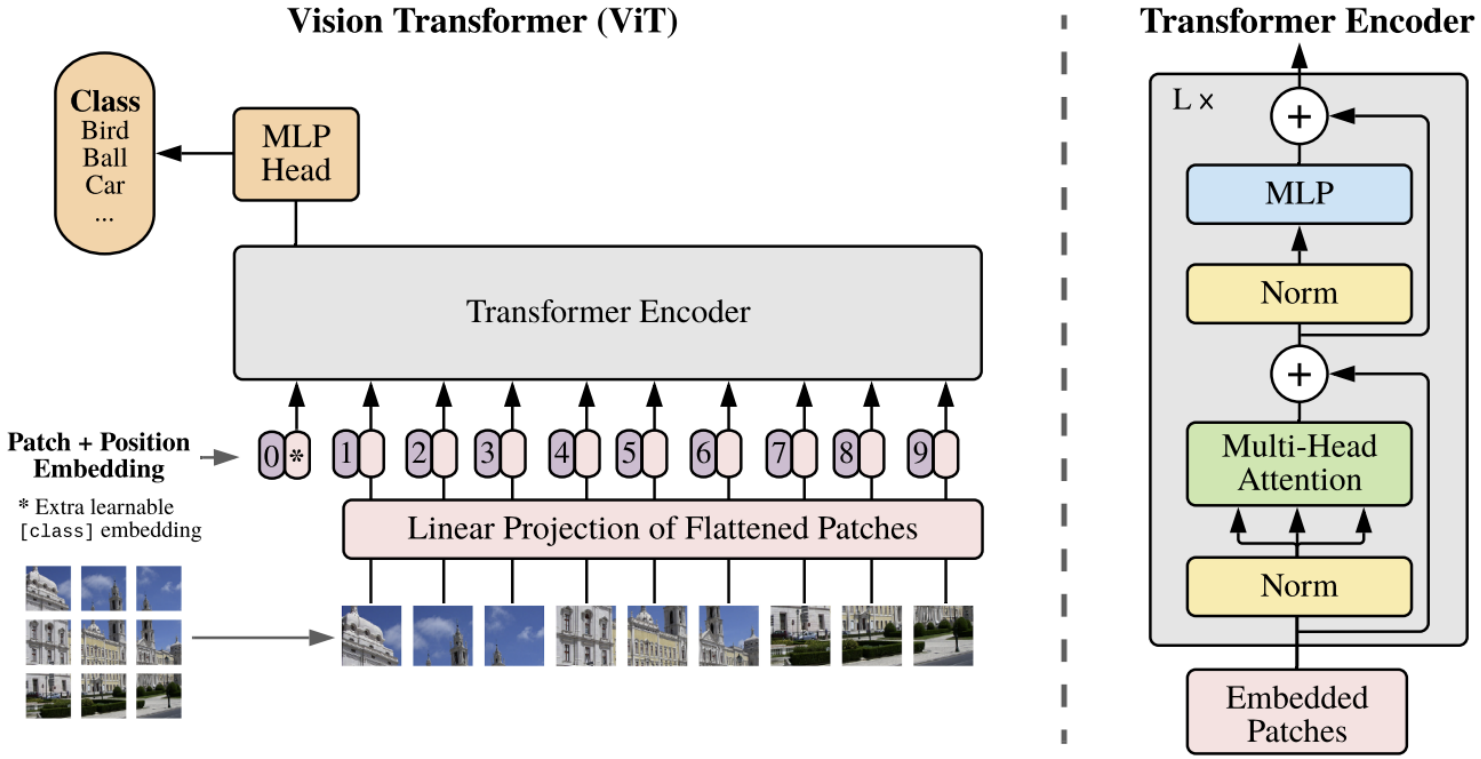
(b) Transformer Backbone (Supervised)

(c) Transformer Backbone (Self-supervised)

Nie et al, “A TIME SERIES IS WORTH 64 WORDS: LONG-TERM FORECASTING WITH TRANSFORMERS”, ICLR 2023.

Transformer in 2D data

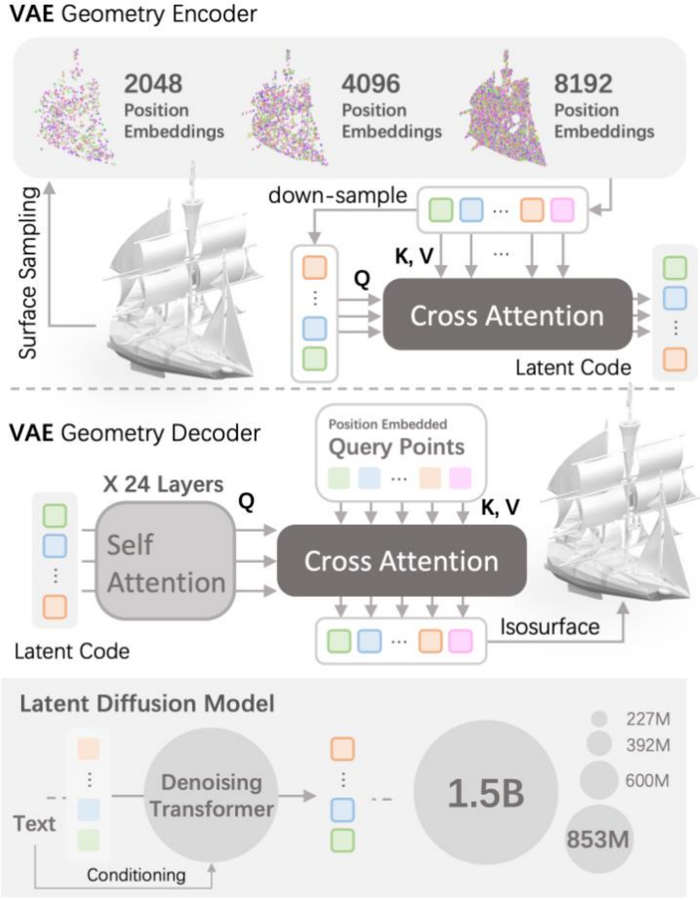
- Vision Transformer (ViT, 10/2020, 50577 citations)
- SOTA performance on ImageNet-1K image classfication

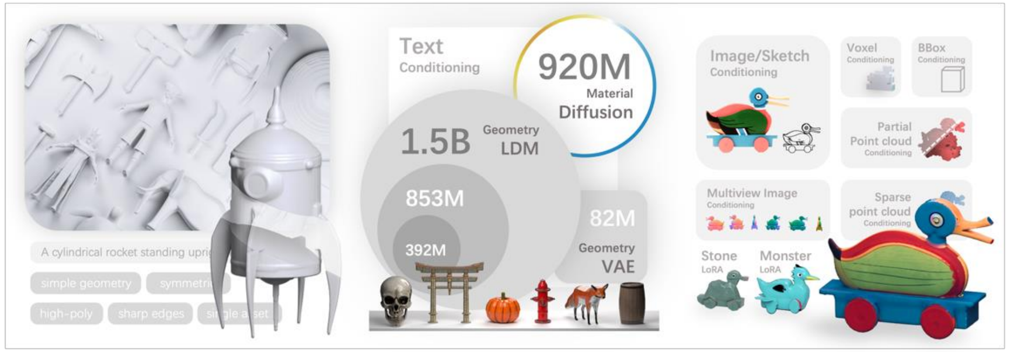


Dosovitskiy et al, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”. ICLR2021

Transformer in 3D data

- CLAY: 3D asset generation model





Zhang L, et al. CLAY: A Controllable Large-scale Generative Model for Creating High-quality 3D Assets. SIGGRAPH 2024. (Best paper honorable mention)

Is attention all we need for real-world AI?

- Attention is all you need. NeurIPS 2017, 146363 citations

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

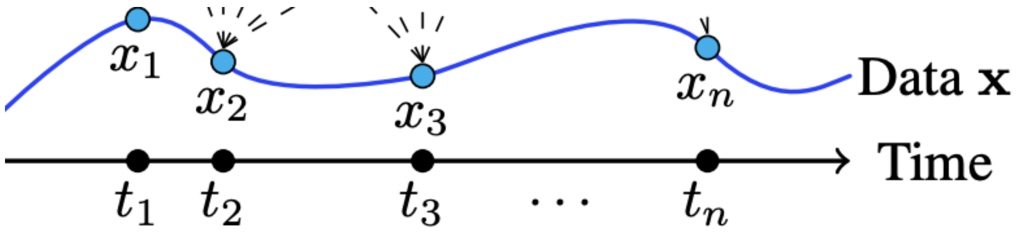
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Is attention all we need for real-world AI?

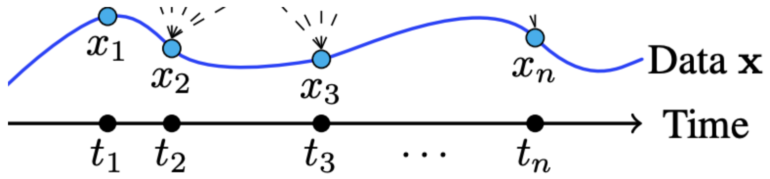
- Limitations of Transformer architecture
  - Huge time and space complexity
  - Little inductive bias and data hungry
  - Discrete time modeling only



Kidger, et al. 2020

Transformer in continuous time space

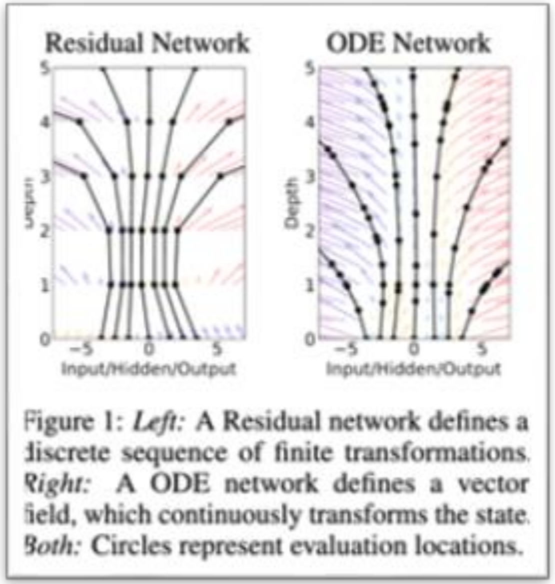
- Continuous-time Transformer (ContiFormer)
  - Empowering Transformer models with continuous-time modeling



Kidger, et al. 2020

Model	Continuous Modeling	Parallelization	Relation Modeling
Transformer[2]	☹️	😊	😊
Neural ODE[3]	😊	☹️	☹️
ContiFormer	😊	😊	😊

Comparison b/w sequence models



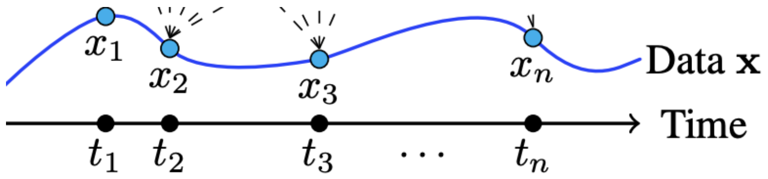
$$\frac{d\mathbf{h}(t)}{dt} = f(\mathbf{h}(t), t, \theta)$$

NeuralODE. 2018
NeurIPS 2018 Best Paper

Chen Y, Ren K, et al. "Contiformer: Continuous-time transformer for irregular time series modeling." Advances in Neural Information Processing Systems (2023).

Transformer in continuous time space

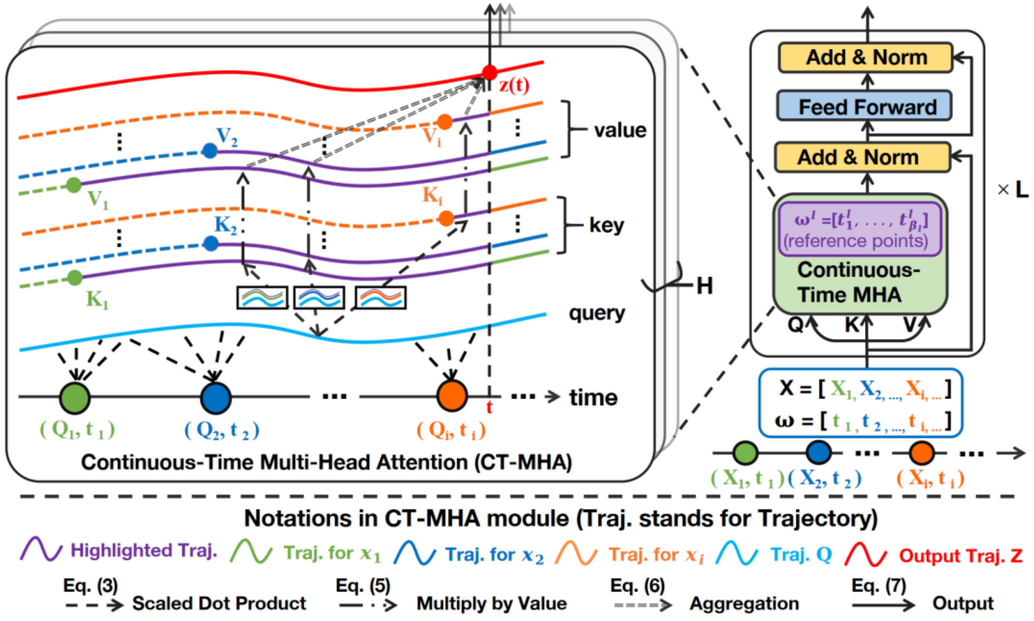
- Continuous-time Transformer (ContiFormer)
  - Empowering Transformer models with continuous-time modeling



Kidger, et al. 2020

Model	Continuous Modeling	Parallelization	Relation Modeling
Transformer[2]	🙁	😊	😊
Neural ODE[3]	😊	🙁	🙁
ContiFormer	😊	😊	😊

Comparison b/w sequence models



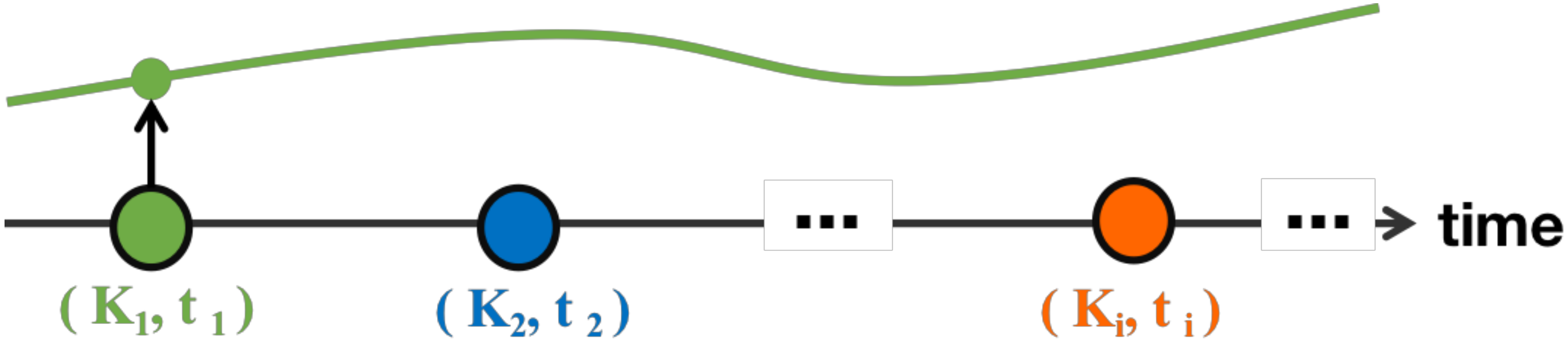
Contiformer. 2023

Chen Y, Ren K, et al. "Contiformer: Continuous-time transformer for irregular time series modeling." Advances in Neural Information Processing Systems (2023).

Model Details

- **Continuous Dynamics from Observations.** We first employ ordinary differential equations (ODEs) to define the latent trajectories for each observation.

$$
\begin{aligned}
\mathbf{k}_i(t_i) = K_i \,, \quad \mathbf{k}_i(t) = \mathbf{k}_i(t_i) + \int_{t_i}^{t} f\left(\tau, \mathbf{k}_i(\tau); \theta_k\right) \mathrm{d}\tau \,, \\
\mathbf{v}_i(t_i) = V_i \,, \quad \mathbf{v}_i(t) = \mathbf{v}_i(t_i) + \int_{t_i}^{t} f\left(\tau, \mathbf{v}_i(\tau); \theta_v\right) \mathrm{d}\tau \,,
\end{aligned}
\tag{1}
$$



Chen Y, Ren K, et al. "Contiformer: Continuous-time transformer for irregular time series modeling." Advances in Neural Information Processing Systems (2023).

Model Details

- **Scaled Dot Product.** Given two real functions $f(x)$ and $g(x)$, the inner product of two functions in a closed interval $[a, b]$ is defined as.

$$\langle f, g \rangle = \int_a^b f(x) \cdot g(x) \mathrm{d}x \ . \tag{2}$$

- Intuitively, it can be thought of as a way of quantifying how much the two functions “align” with each other over the interval. Inspired by the formulation of inner product in continuous time. Therefore, we model the evolving relationship between the i-th sample and the dynamic system at time point t as the inner product of q and k_i.

$$\boldsymbol{\alpha}_i(t) = \frac{\int_{t_i}^{t} \mathbf{q}(\tau) \cdot \mathbf{k}_i(\tau)^\top \mathrm{d}\tau}{t - t_i} \ . \tag{3}$$

Chen Y, Ren K, et al. "Contiformer: Continuous-time transformer for irregular time series modeling." Advances in Neural Information Processing Systems (2023).

Experiments: modelling continuous-time function

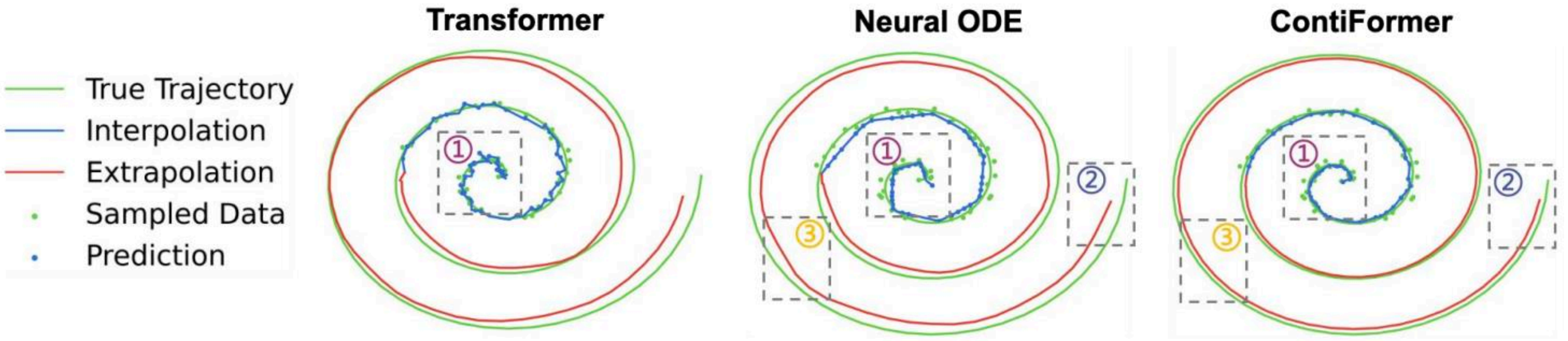


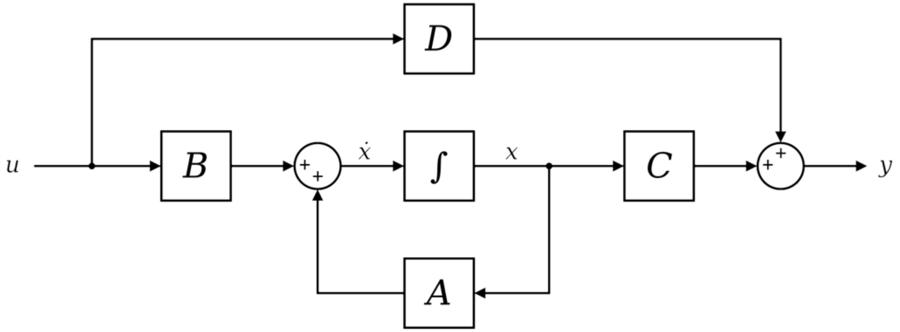
Figure 2: Interpolation and extrapolation of spirals with irregularly-samples time points by Transformer, Neural ODE and our model.

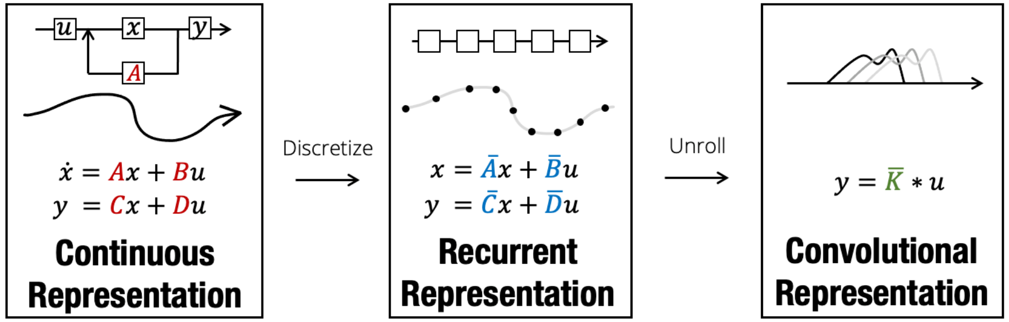
- Transformer fails to output a smooth and continuous function approximation given the noisy observations (①).
- ContiFormer excels at retaining long-term information (②).
- Neural ODE is prone to cumulative errors (③).

Chen Y, Ren K, et al. "Contiformer: Continuous-time transformer for irregular time series modeling." Advances in Neural Information Processing Systems (2023).

State-space model

- State-Space Models (S4, Mamba, etc.)
  - Represent dynamic systems using state variables, capturing system behavior through first-order differential or difference equations.
  - Applications
    - Widely utilized in control theory, econometrics, neuroscience, and machine learning for modeling time-series data and dynamic processes.





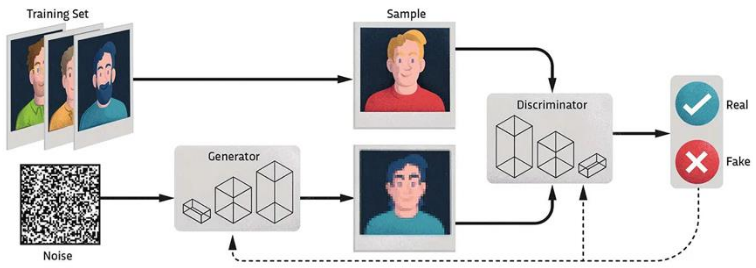
Gu, et al. "Efficiently Modeling Long Sequences with Structured State Spaces." In *International Conference on Learning Representations. 2022. (Outstanding paper honorable mention)*

Advanced topics

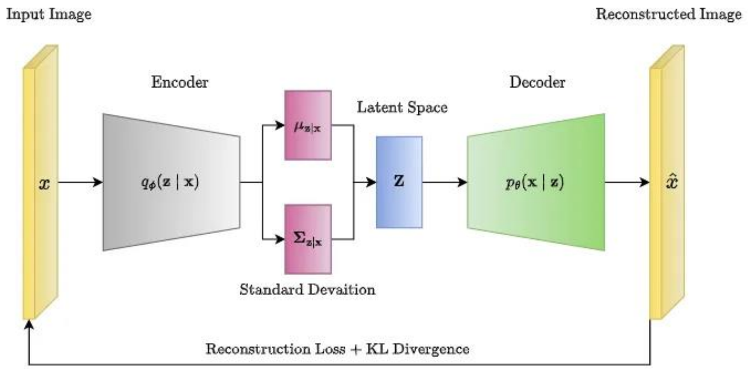
- Advanced deep learning
  - Transformer
  - Generative modeling
  - Foundation model
- Advanced reinforcement learning
  - Real-world reinforcement learning
  - AI agents
- Responsible AI

Generative model

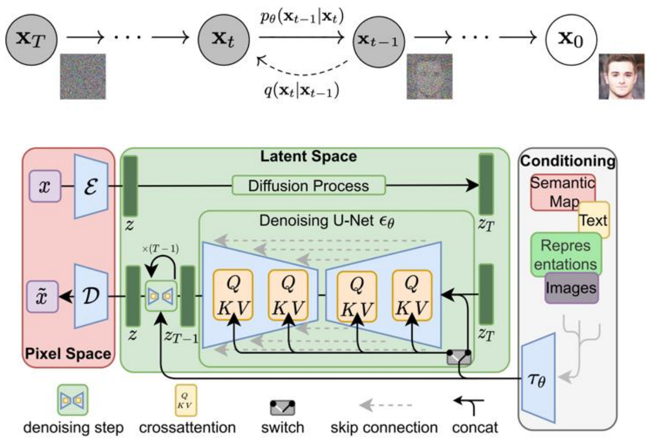
- Transferring from discriminative model to generative model
  - Discriminative model: $p(y|x;\theta)$
  - Generative model: $p(x, y; \theta)$ or even without label $p(x;\theta)$



Generative adversarial network (GAN)
Goodfellow et al. 2014.
NeurIPS Test of Time Award



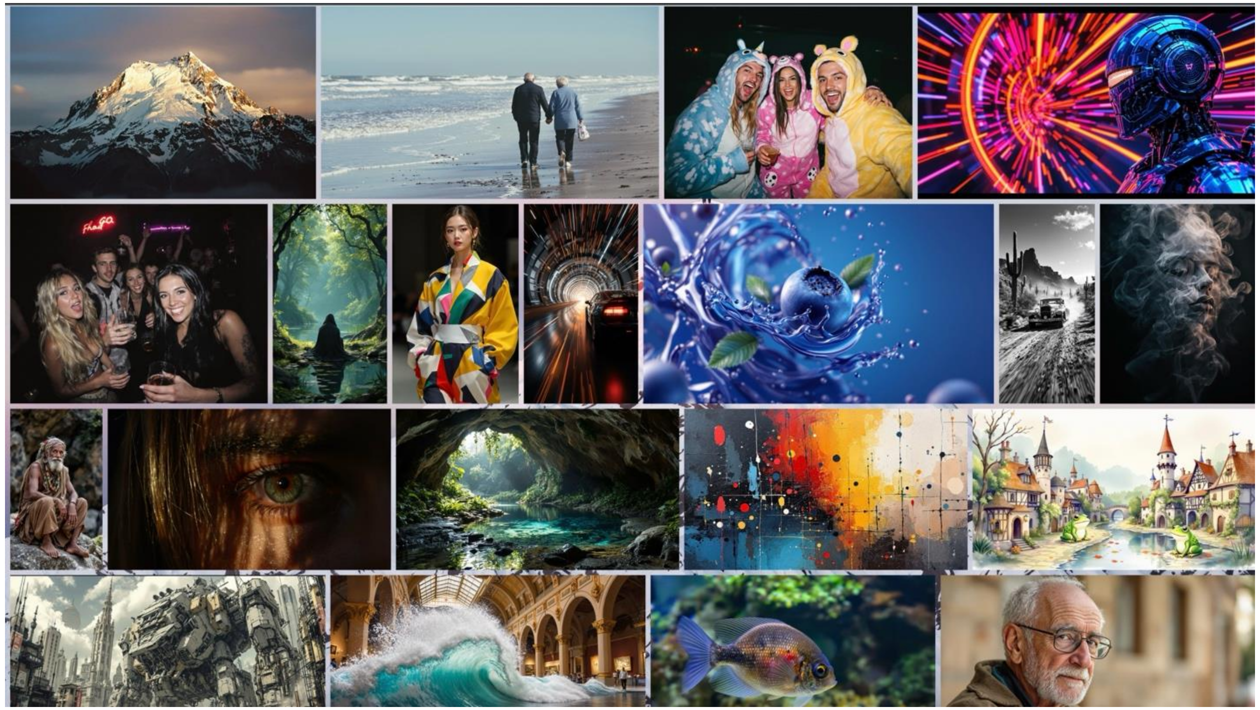
Autoencoder and Variational Autoencoder
Kingma et al. 2014.
ICLR Test of Time Award



Diffusion models
Ho et al. 2020.

Text-to-image Generative Models

Stable Diffusion 3.5[1] is creative



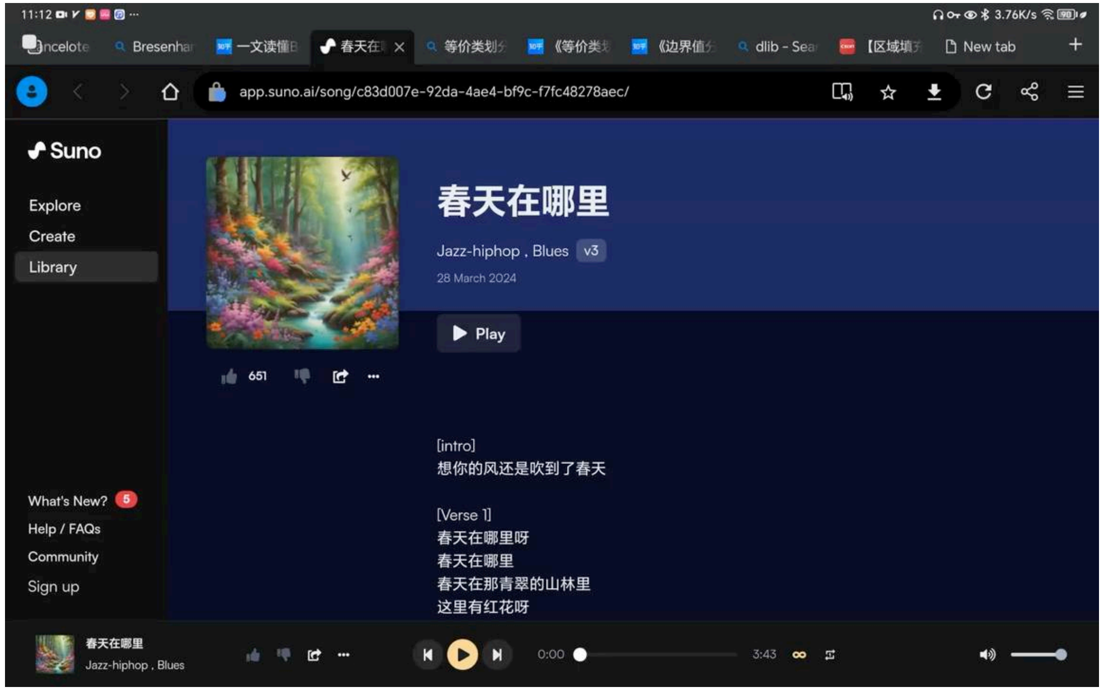
Flux 1.1 Pro[2] can generate realistic photos

Both Stable Diffusion 3.5 and Flux 1.1 Pro are Open Source and available in huggingface.

[1] Stable Diffusion 3.5 Large, https://huggingface.co/stabilityai/stable-diffusion-3.5-large
[2] Flux 1.1 Pro, https://blackforestlabs.ai/ultra-home

Text-to-music Generative Models

DiffSinger[1] – AI孙燕姿



Suno AI[2]

DiffSinger is Open Source in github. Suno AI is Closed Source, but available at suno.com

[1] DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism (AAAI 2022)
[2] Suno AI, https://suno.com/

Video Generative Models

Kling (快手可灵) can generate Lifelike Large Motions Videos and Minute-level Long Videos up to 2 minutes length.

OpenAI Sora demo

[1] 快手可灵, https://kling.kuaishou.com/en
[2] OpenAI Sora, https://openai.com/index/sora-is-here/

3D Generative Models

3D world created by Worldlabs[1] and Fei-Fei Li.

[1] Worldlabs, https://www.worldlabs.ai/

Generative model

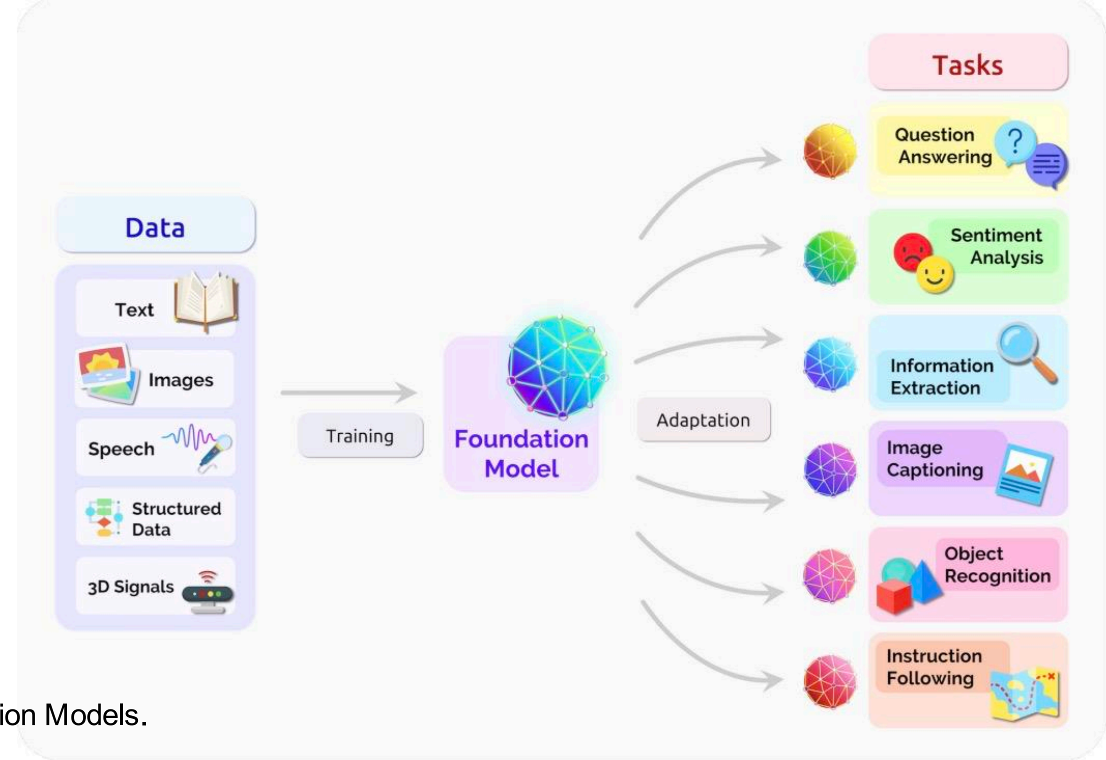
- Transferring from discriminative model to generative model
  - Discriminative model: $p(y|x;\theta)$
  - Generative model: $p(x,y;\theta)$ or even without label $p(x;\theta)$

- Now that generative model captures the joint distribution, we can transform the different tasks into posterior form

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Foundation model

- Definition
  - **Large-scale pre-trained models** designed to serve as a general-purpose platform for a wide range of downstream tasks.



[1] On the Opportunities and Risks of Foundation Models.

Foundation model

- Definition
  - **Large-scale pre-trained models** designed to serve as a general-purpose platform for a wide range of downstream tasks.
- Key Characteristics
  - Scale: Trained on massive datasets and often consist of billions of parameters.
  - Generalization: Capable of handling diverse tasks, including text generation, image synthesis, and more.
  - Transferability: Fine-tuned to specific tasks with minimal labeled data.
  - Versatility: Used in diverse domains such as language, vision, audio, and multimodal tasks.

Foundation model

What is this a picture of?

This is an image of a customized card resembling a Pokémon trading card. The card features an adorable sleeping Pekingese dog wrapped in a blanket, with the name "Pekechuu" at the top. The card includes two moves both named "Nap" with a value of 0. The design mimics the layout and style of a traditional Pokémon card, including the card number, type, and various stats, though some of the text appears to be in a playful or nonsensical format.

User

User
Do you know who drew this painting?

LLaVA
The painting depicts a woman, commonly believed to be Mona Lisa, the famous artwork by Leonardo da Vinci. It is a portrait painting that showcases the woman's enigmatic smile and has become one of the most famous and iconic art pieces in the world. The original work is displayed in the Louvre Museum in Paris, and it is known for its intricate details, use of oil paint, and the artist's innovative techniques that contributed to its enduring appeal and mystery.

Both images and texts inputs are accepted in GPT-4o [1].

There are also excellent Open Source Multimodality models like LLaVA [2].

[1] GPT-4o, https://openai.com/index/hello-gpt-4o/
[2] LLaVA Visual Instruction Tuning. (NIPS2023 Oral)

Foundation model

- Examples
  - Language Models: GPT (OpenAI), BERT (Google), LLaMA (Meta).
  - Vision Models: CLIP (OpenAI), SAM (Meta).
  - Multimodal Models: DALL-E (OpenAI), Flamingo (DeepMind), Sora (OpenAI), GPT4V (OpenAI).

- Advantages
  - Reduce the need for task-specific training.
  - Enable rapid development of AI solutions.
  - Foster cross-domain innovation through unified architecture.

Are (pretrained) foundation models enough for AI?

- Pretraining has been a tremendous surge in recent years.

Pretrained models on *ImageNet*
Cadene/pretrained-models, 45 models

https://github.com/Cadene/pretrained-models.pytorch

README.md

Evaluation on imagenet

Accuracy on validation set (single model)

Results were obtained using (center cropped) images of the same size t

Model	Version	Acc@1	Acc@5
PNASNet-5-Large	Tensorflow	82.858	96.182
PNASNet-5-Large	Our porting	82.736	95.992
NASNet-A-Large	Tensorflow	82.693	96.163
NASNet-A-Large	Our porting	82.566	96.086
SENet154	Caffe	81.32	95.53
SENet154	Our porting	81.304	95.498
PolyNet	Caffe	81.29	95.75
PolyNet	Our porting	81.002	95.624
InceptionResNetV2	Tensorflow	80.4	95.3
InceptionV4	Tensorflow	80.2	95.3

PyTorch (pretrained) Image Models
rwightman/pytorch-image-models, 79 models

rwightman/pytorch-image-mod

https://github.com/rwightman/pytorch-image-...

README.md

Introduction

PyTorch Image Models (`timm`) is a collection of image models, layers, utilities, optimizers, schedulers, data-loaders / augmentations, and reference training / validation scripts that aim to pull together a wide variety of SOTA models with ability to reproduce ImageNet training results.

The work of many others is present here. I've tried to make sure all source material is acknowledged via links to github, arxiv papers, etc in the README, documentation, and code docstrings. Please let me know if I missed anything.

Models

All model architecture families include variants with pretrained weights. There are specific model variants without any weights, it is NOT a bug. Help training new or better weights is always appreciated. Here are some example training hparams to get you started.

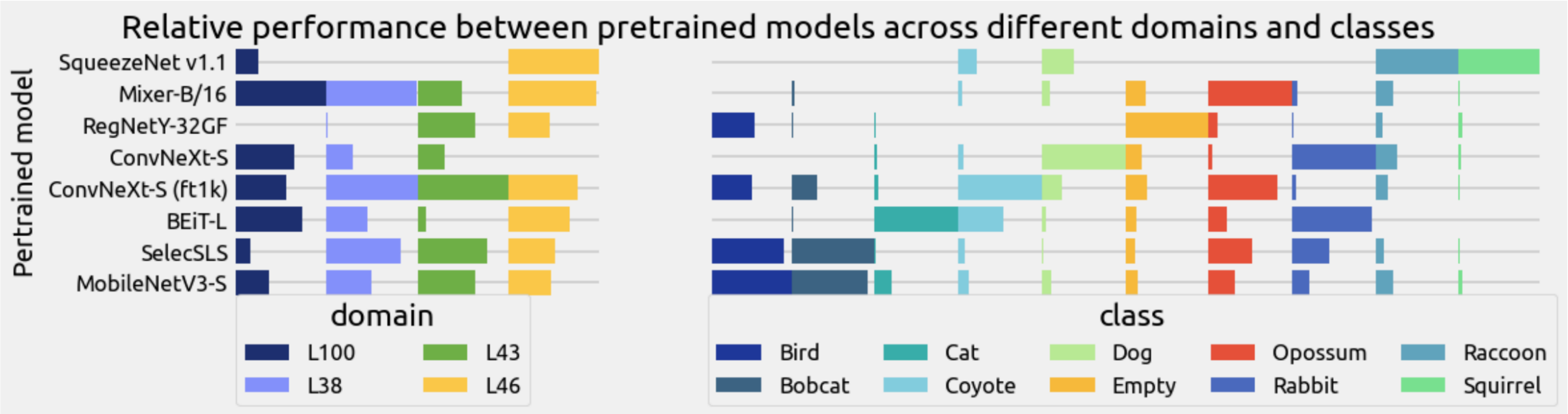
A full version of the list below with source links can be found in the documentation.

- Aggregating Nested Transformers - https://arxiv.org/abs/2105.12723
- BEiT - https://arxiv.org/abs/2106.08254
- Big Transfer ResNetV2 (BiT) - https://arxiv.org/abs/1912.11370
- Bottleneck Transformers - https://arxiv.org/abs/2101.11605
- CaiT (Class-Attention in Image Transformers) - https://arxiv.org/abs/2103.17239
- CoaT (Co-Scale Conv-Attentional Image Transformers) - https://arxiv.org/abs/2104.06399
- ConvNeXt - https://arxiv.org/abs/2201.03545

Li Z, Ren K, et al. "Simple: Specialized model-sample matching for domain generalization." *The Eleventh International Conference on Learning Representations*. 2023.

Are (pretrained) foundation models enough for AI?

- No free lunch for pretrained models in generalization



- ***Takeaway 1***: Pretrained models possess decent generalization ability for some OOD samples.
- ***Takeaway 2***: No dominant pretrained models across unseen domains.
- ***Takeaway 3***: Pretrained models exhibit more diverse performance at finer-grained levels.

Li Z, Ren K, et al. "Simple: Specialized model-sample matching for domain generalization." *The Eleventh International Conference on Learning Representations*. 2023.

Are (pretrained) foundation models enough for AI?

- Shift in test distribution may help or hurt generalization

$$\underbrace{E_g}_{\text{OOD error}} = \underbrace{E_g^{matched}}_{\text{IID error}} + \kappa \bar{\mathbf{a}}^T (P\mathbf{\Lambda} + \kappa \mathbf{I})^{-1} \underbrace{\mathcal{O}'}_{\substack{\text{alignment between} \\ \text{the test distribution and the model}}} (P\mathbf{\Lambda} + \kappa \mathbf{I})^{-1} \bar{\mathbf{a}}$$

- Should not the focus be more on matching the pretrained model and test distributions based on their fitness?

Li Z, Ren K, et al. "Simple: Specialized model-sample matching for domain generalization." *The Eleventh International Conference on Learning Representations*. 2023.

Are (pretrained) foundation models enough for AI?

- Shift in test distribution may help or hurt generalization

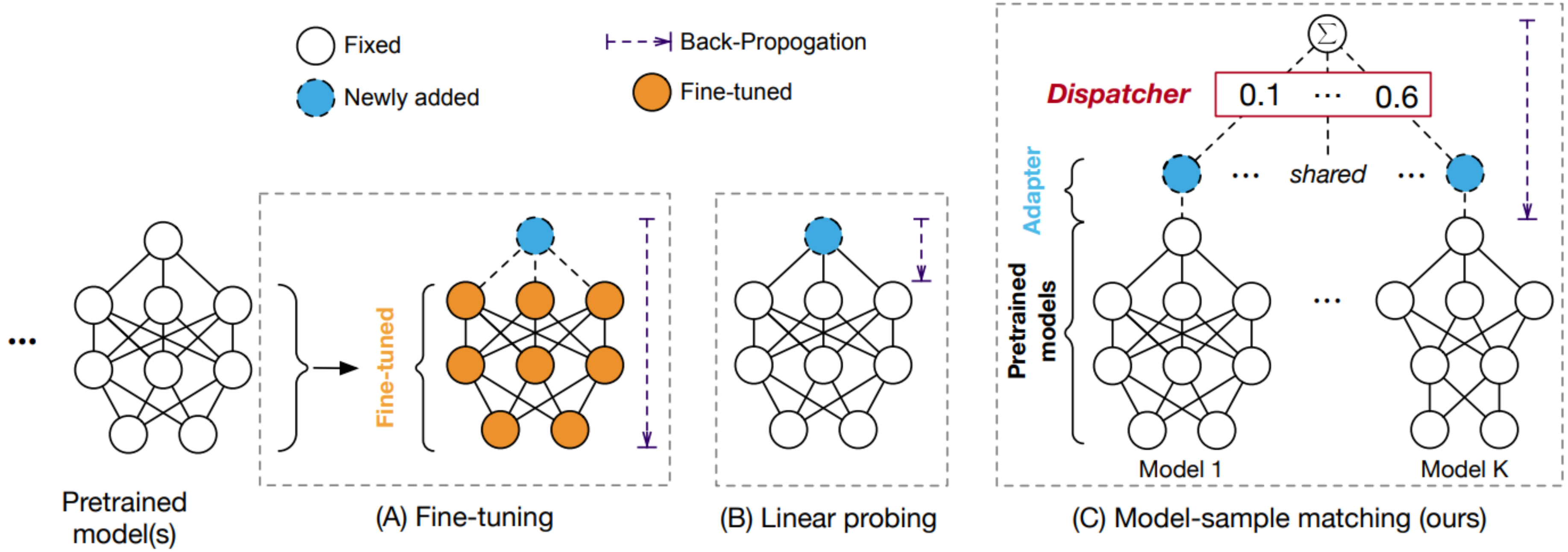


Figure 2: Different training paradigms in DG.

Li Z, Ren K, et al. "Simple: Specialized model-sample matching for domain generalization." *The Eleventh International Conference on Learning Representations*. 2023.

Are (pretrained) foundation models enough for AI?

- **∑[samller pretrained models] > a large one**
  - More effective
    - **Gains up to 3.9%** on domain generalization benchmark
    - **Up to 12.2%** on single dataset
  - More efficient
    - Training cost: **1000 times training speedup** compared to conventional methods
    - Inference cost: New SOTA results with **higher inference efficiency**, since we only select a small set of the pretrained models for inference

Li Z, Ren K, et al. "Simple: Specialized model-sample matching for domain generalization." *The Eleventh International Conference on Learning Representations*. 2023.

Advanced topics

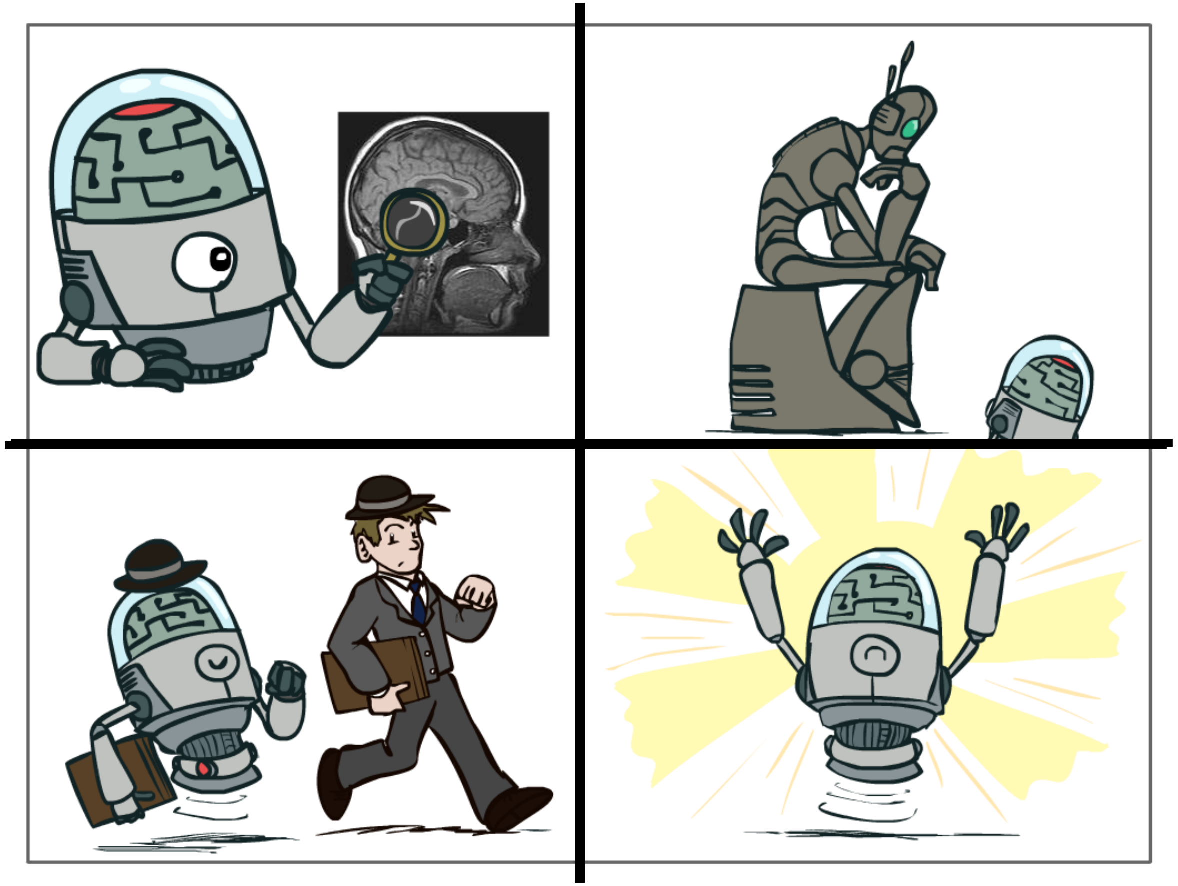
- Advanced deep learning
  - Transformer
  - Generative modeling
  - Foundation model
- Advanced reinforcement learning
  - Real-world reinforcement learning
  - AI agents
- Responsible AI

Recap of rational AI agents

Should we make machines that…

Think like people?

Think rationally?

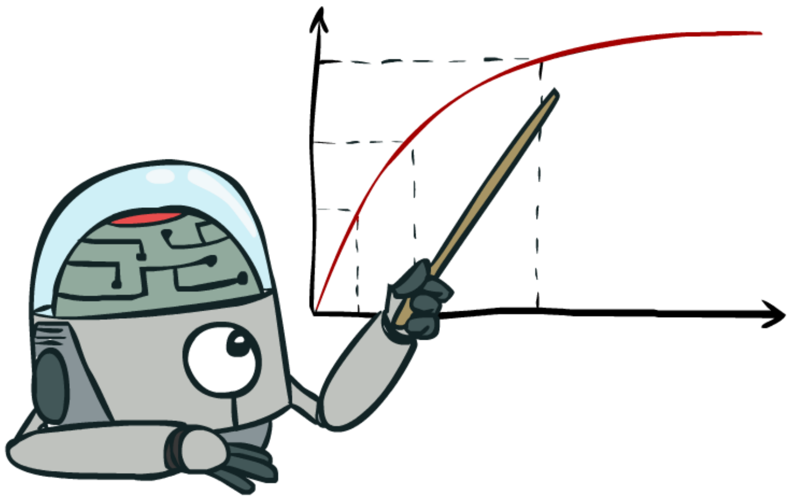


Act like people?

Act rationally?

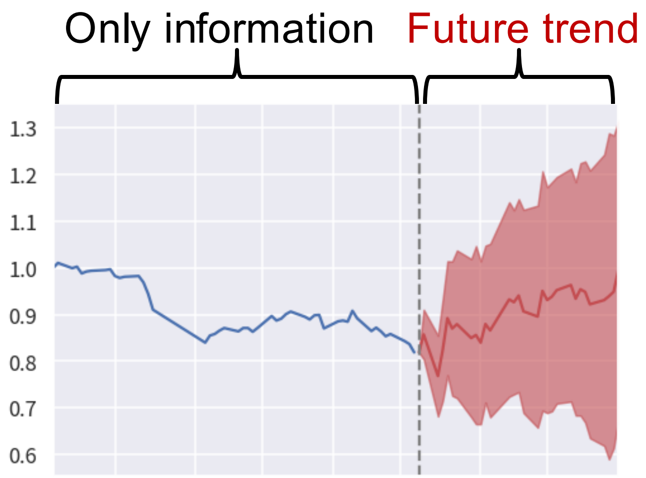
Recap of rational AI agents: Rational Decisions

- We'll use the term **rational** in a very specific, technical way:
  - Rational: *maximally achieving pre-defined goals*
  - Goals are expressed in terms of the **utility** of outcomes
  - World is uncertain, so we'll use **expected** utility
  - Being rational means acting to **maximize your expected utility**



Decision making agent: beyond prediction

- Go beyond simply prediction-based decision making
  - Given the observation x_t,
  - the model-based methods often *assume* a model (function) with some prediction target as $\hat{p}_t = f(x_t)$,
  - Then they derive the decision *upon the intermediate prediction* as $a_t = g(o_t)$.



How about directly optimize the decision making?

Reinforcement learning

$\boldsymbol{x}_t \rightarrow \hat{p}_t \rightarrow a_t$

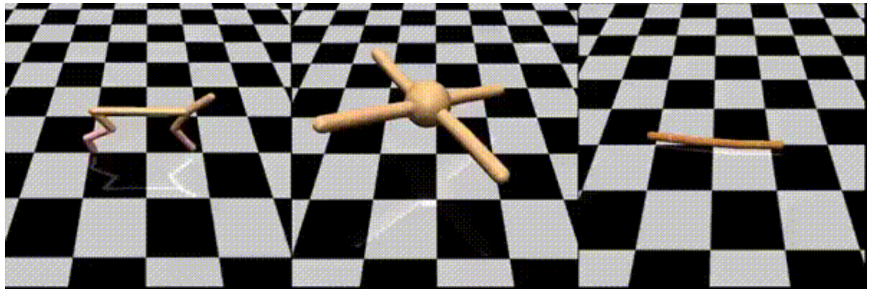
Traditional model-based solution

Sub-optimality

The decision making is not optimized directly, fail to manage noise and easy to overfit.

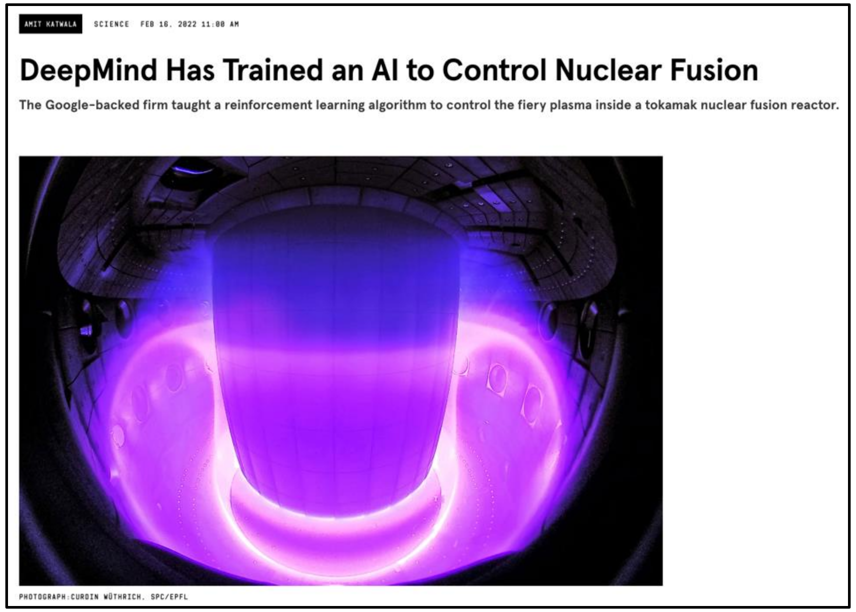
Amazing achievements of RL

Logistics system



MuJoCo Robot Control

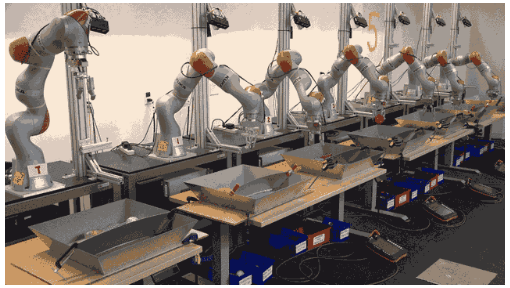
Video Games


AMIT KATWALA SCIENCE FEB 16, 2022 11:00 AM

DeepMind Has Trained an AI to Control Nuclear Fusion

The Google-backed firm taught a reinforcement learning algorithm to control the fiery plasma inside a tokamak nuclear fusion reactor.

PHOTOGRAPH:CURDIN WÜTHRICH, SPC/EPFL



Industrial production

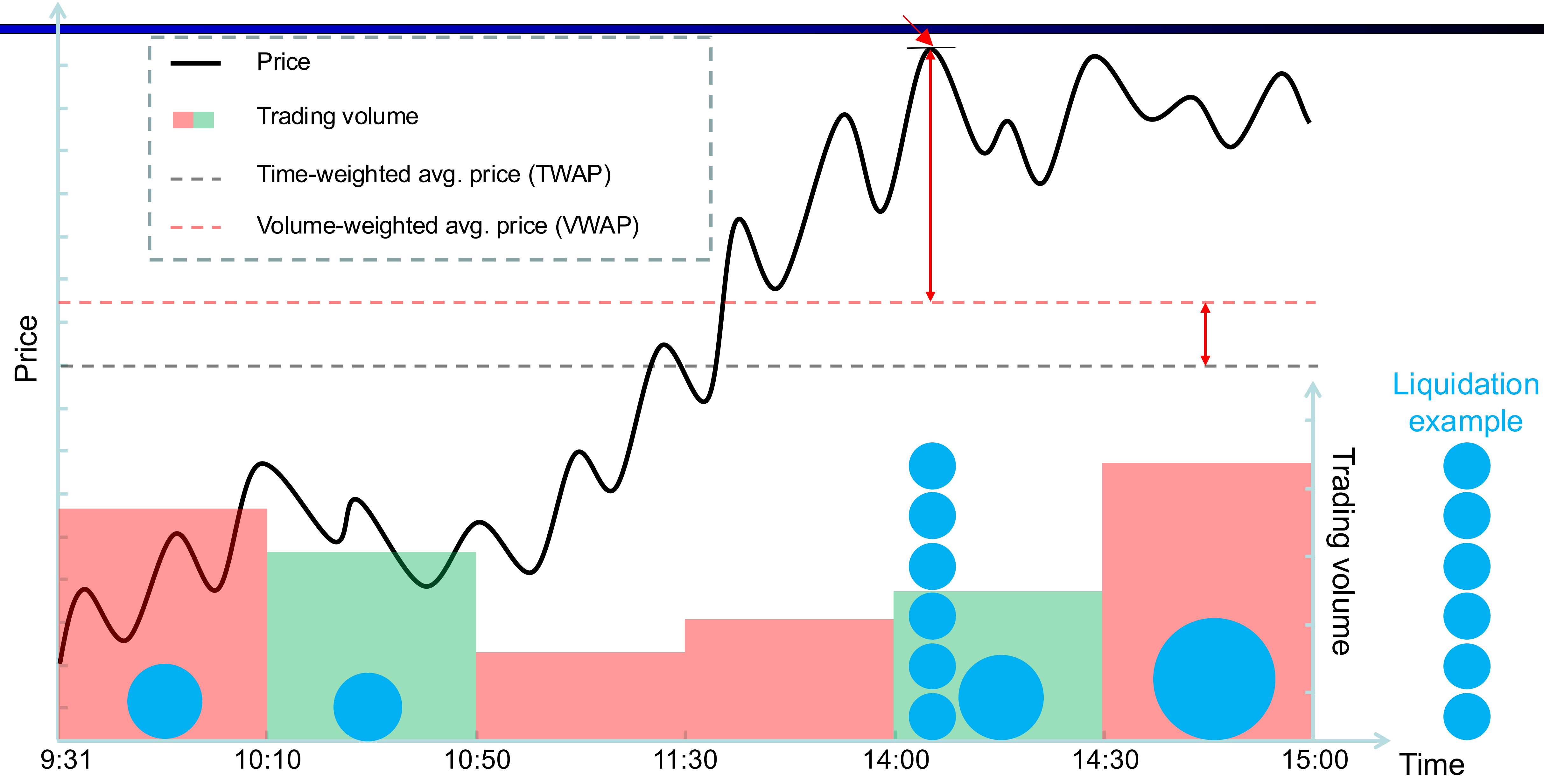
Recommendation and ads

What’s the killer application of RL (decision AI)

- Answer varies.
- But suitable scenarios may require
  - Infinitely runnable evaluation
  - Sandbox-like environments
  - Easy-to-obtain feedbacks
- Typical scenarios
  - Game
  - Finance
  - Chatbot
  - Anything else with reasonable simulators

Example: financial trading

Background of financial **order execution**



Order execution task

- Order execution is
  - to trade (buy/sell) the specific number of stock shares, at better price,
  - during a trading time horizon $[0, T]$.

  - A sequential decision-making procedure since decisions are correlated.

$$\textit{min}/\textit{max} \sum_{t=0}^{T} (p_t q_t)$$

$$\text{s.t. } \sum_{t=0}^{T} q_t = Q$$

Trading as a Markov decision process

- Take order execution as a direct sequential decision optimization.
- Markov decision process assumption

Notation	Markov decision process	Information
$\boldsymbol{s}_t$	State	Private: left order to trade, timestep
		Public: market price & volume information
a_t	Action	The proportion of order to trade at the next timestep
$r_t(s_t, a_t)$	Reward	Weighted price advantage $\widehat{R}_t^+ = \frac{q_{t+1}}{Q} \cdot \frac{p_{t+1} - \tilde{p}}{\tilde{p}} = a_t(\frac{p_{t+1}}{\tilde{p}} - 1)$ TWAP
		Large sub-order penalty $\bar{R}_t^- = -\alpha(a_t)^2$
$P(s_t, a_t)$	Probability transition	$P(\boldsymbol{s}_t \mid \boldsymbol{s}_0, a_0, \dots, \boldsymbol{s}_{t-1}, a_{t-1}) = P(\boldsymbol{s}_t \mid \boldsymbol{s}_{t-1}, a_{t-1})$
γ	Discount rate	$\gamma = 1.0$

[1] Fang Y, Ren K, et al. "Universal trading for order execution with oracle policy distillation." In Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

Evaluation

- Metric
  - **PA:** price advantage (‱) to the average market price (TWAP)
  - **Reward** including negative market impacts

Category	Strategy	Reward($\times 10^{-2}$)	PA
financial model-based	TWAP (Bertsimas et al. 1998)	-0.42	0
	AC (Almgren et al. 2001)	-1.45	2.33
	VWAP (Kakade et al. 2004)	-0.30	0.32
learning-based	DDQN (Ning et al. 2018)	2.91	4.13
	PPO (Lin et al. 2020)	1.32	2.52
	OPD^S (pure student)	3.24	5.19
	OPD (our proposed)	**3.36***	**6.17***

Table 2: Performance comparison; the higher, the better.

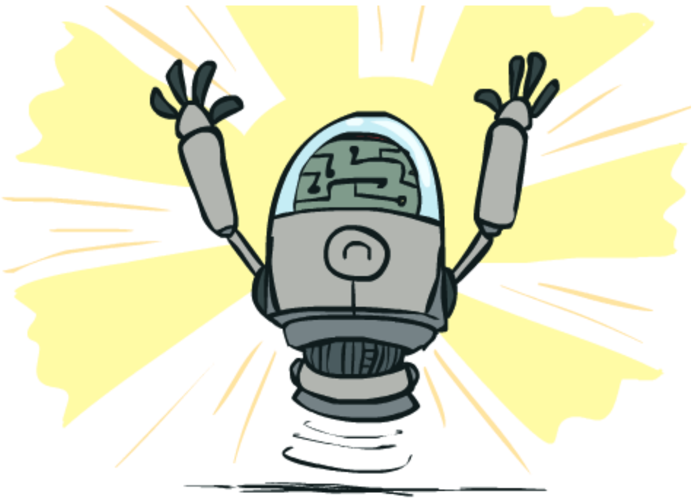
PA	Annual return rate
1.0	0.55%
5.0	2.75%

A little bit higher than fixed term deposit.

[1] Fang Y, Ren K, et al. "Universal trading for order execution with oracle policy distillation." In Proceedings of the AAAI Conference on Artificial Intelligence, 2021.

Wait, one more thing …

- Do AI agents really realize what they are doing?
  - Or do they just optimize the cumulative rewards without cognition?



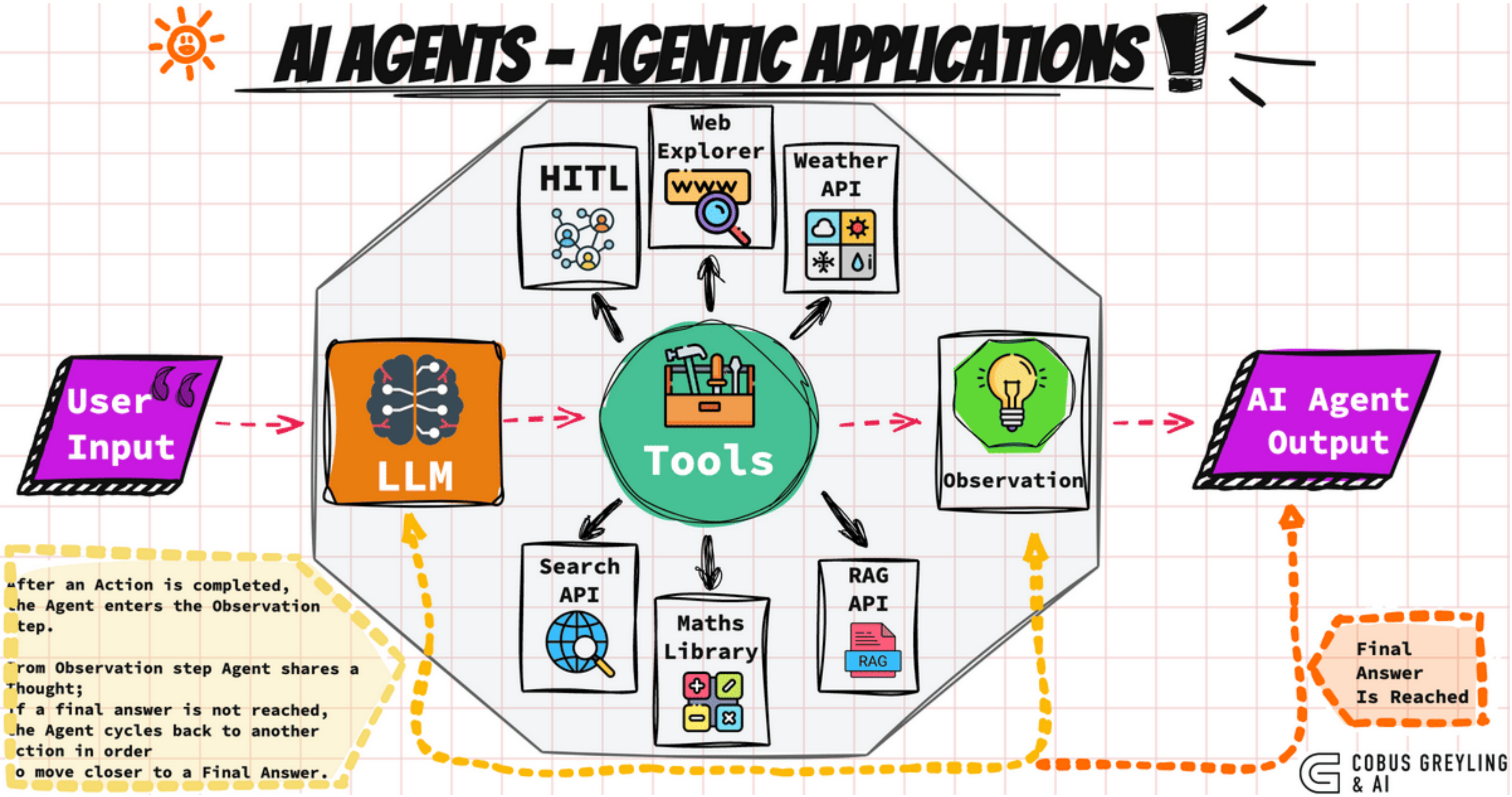
Wait, one more thing …

- Do AI agents really realize what they are doing?
  - Or do they just optimize the cumulative rewards without cognition?

- On one hand,
  - AI should know what are really underlying the task it's doing

Rational and reasoning agents

- Solving complex real-world tasks via reasoning agents



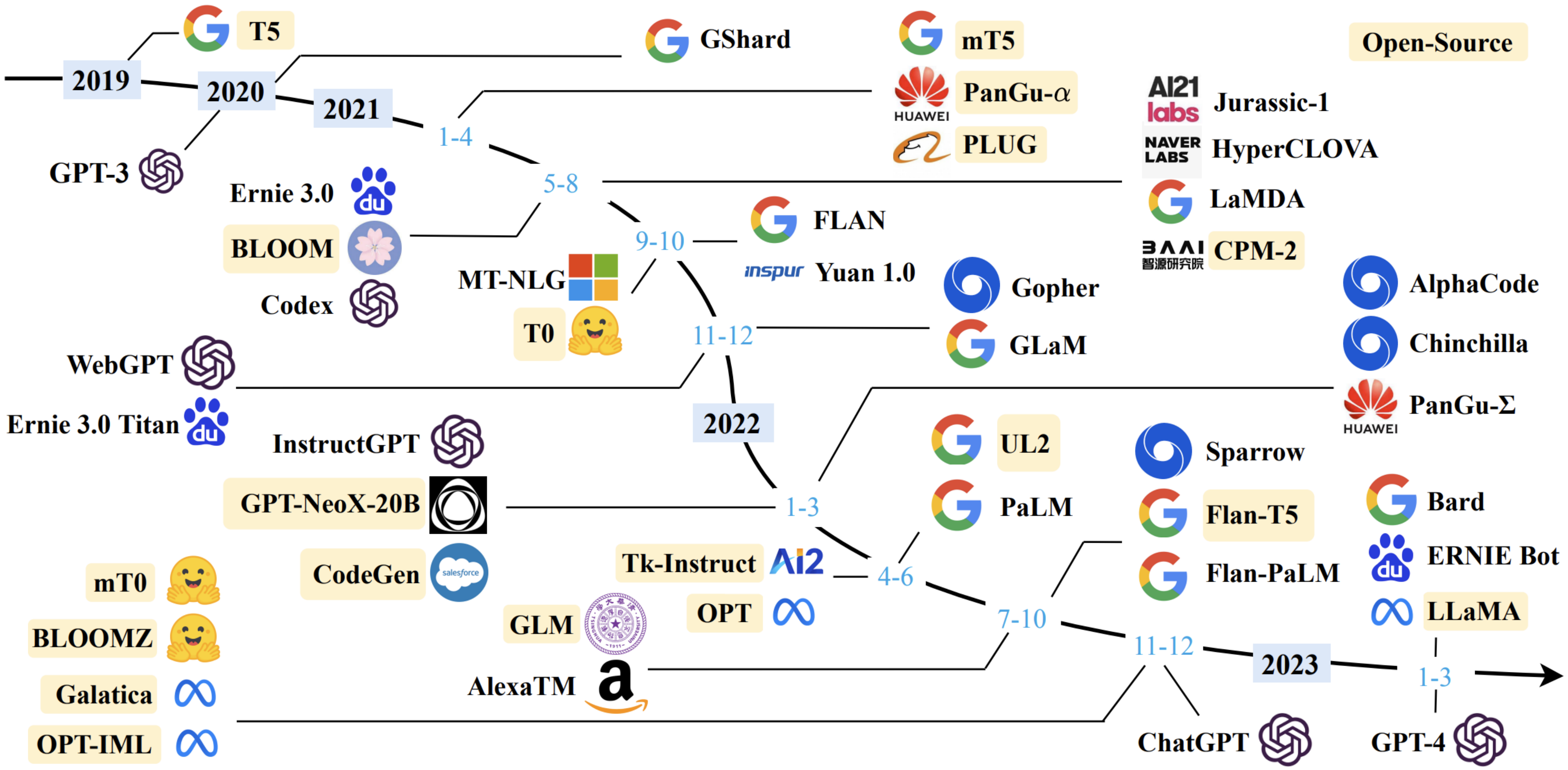
Rational and reasoning agents

- Solving complex real-world tasks via reasoning agents

OpenAI DevDay 2024

Rational and reasoning agents

- LLM-based decision agents



Data science agents

- For data scientists, they should handle a bunch of libraries.

`torch.optim.Adam(lr=?, betas=?)`

sklearn.ensemble: Ensemble Methods

The sklearn.ensemble module includes ensemble-based methods for classification, regression and anomaly detection.

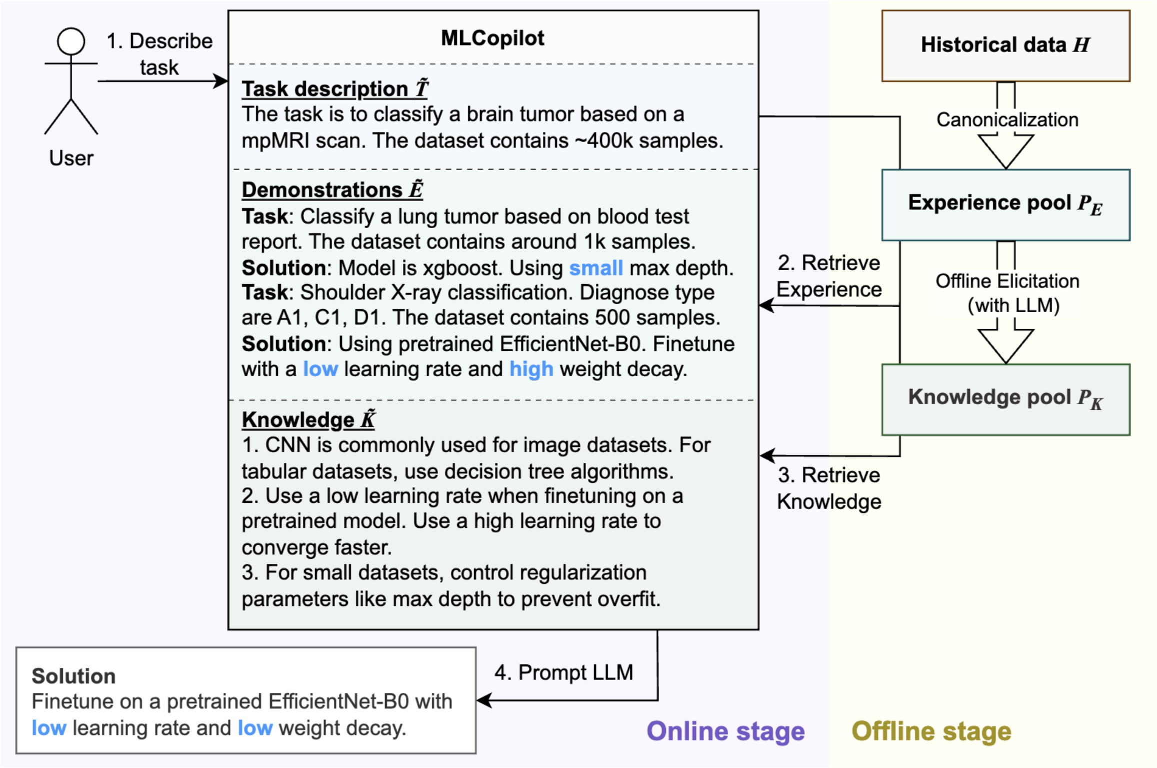
User guide: See the Ensemble methods section for further details.

ensemble.AdaBoostClassifier([estimator, ...])	An AdaBoost classifier.
ensemble.AdaBoostRegressor([estimator, ...])	An AdaBoost regressor.
ensemble.BaggingClassifier([estimator, ...])	A Bagging classifier.
ensemble.BaggingRegressor([estimator, ...])	A Bagging regressor.
ensemble.ExtraTreesClassifier([...])	An extra-trees classifier.
ensemble.ExtraTreesRegressor([n_estimators, ...])	An extra-trees regressor.
ensemble.GradientBoostingClassifier(*[, ...])	Gradient Boosting for classification.
ensemble.GradientBoostingRegressor(*[, ...])	Gradient Boosting for regression.
ensemble.IsolationForest(*[, n_estimators, ...])	Isolation Forest Algorithm.
ensemble.RandomForestClassifier([...])	A random forest classifier.
ensemble.RandomForestRegressor([...])	A random forest regressor.
ensemble.RandomTreesEmbedding([...])	An ensemble of totally random trees.
ensemble.StackingClassifier(estimators[, ...])	Stack of estimators with a final classifier.
ensemble.StackingRegressor(estimators[, ...])	Stack of estimators with a final regressor.
ensemble.VotingClassifier(estimators, *[, ...])	Soft Voting/Majority Rule classifier for unfitted estimators.
ensemble.VotingRegressor(estimators, *[, ...])	Prediction voting regressor for unfitted estimators.
ensemble.HistGradientBoostingRegressor([...])	Histogram-based Gradient Boosting Regression Tree.
ensemble.HistGradientBoostingClassifier([...])	Histogram-based Gradient Boosting Classification Tree.

- Chen N, Zhang Y, Xu J, Ren K, Yang Y. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. IEEE Visualization Conference 2024. (Best Paper Award)
- Zhang L, Zhang Y, Ren K, Li D, Yang Y. MLCopilot: Unleashing the Power of Large Language Models in Solving Machine Learning Tasks. EACL 2024. (Outstanding Paper Award)
- Zhang Y, Jiang Q, Han X, Chen N, Yang Y, Ren K. Benchmarking Data Science Agents. ACL 2024.

Data science agents

- Knowledge extraction and in-context learning (ICL)



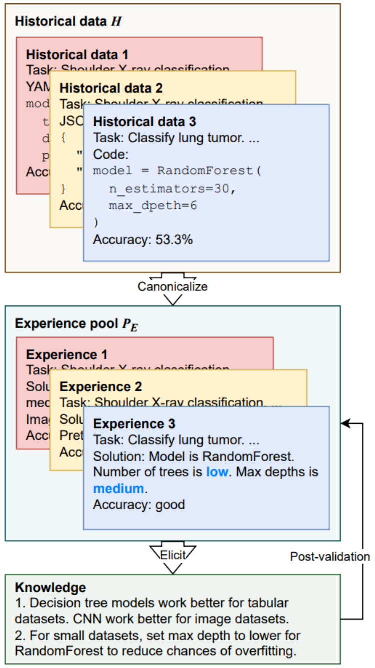


Figure 2: Offline stage: canonicalization, knowledge elicitation.

E Knowledge

All contents in this section are generated by Large Language Models.

E.1 HPO-B

HPO-B contains 16 design spaces. We finalize one set of knowledge for each space.

Space: 5860

1. Generally, datasets with more numeric features require larger alphas and smaller lambdas for better performance.
2. Datasets with a higher ratio of minority to majority class size require smaller alphas and larger lambdas for better performance.
3. Datasets with more features require larger alphas and smaller lambdas for better performance.
4. Datasets with more categorical features require larger alphas and larger lambdas for better performance.

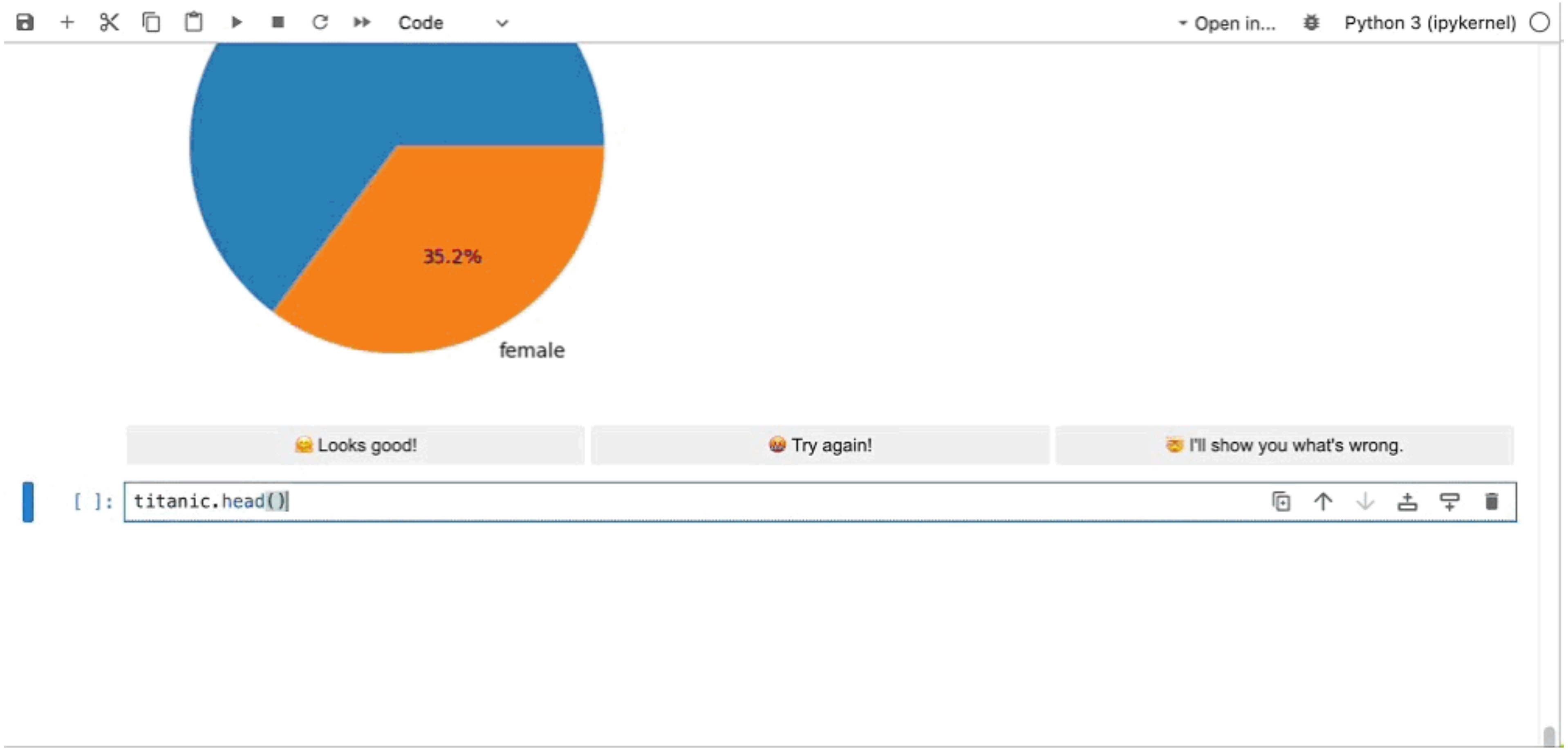
Space: 4796

1. For datasets with a large majority class size and a small minority class size, a larger cp and minbucket size tend to be better hyper-parameter configurations.
2. For datasets with a small majority class size and a large minority class size, a smaller cp and minbucket size tend to be better hyper-parameter configurations.

- Chen N, Zhang Y, Xu J, Ren K, Yang Y. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. IEEE Visualization Conference 2024. (Best Paper Award)
- Zhang L, Zhang Y, Ren K, Li D, Yang Y. MLCopilot: Unleashing the Power of Large Language Models in Solving Machine Learning Tasks. EACL 2024. (Outstanding Paper Award)
- Zhang Y, Jiang Q, Han X, Chen N, Yang Y, Ren K. Benchmarking Data Science Agents. ACL 2024.

Data science agents

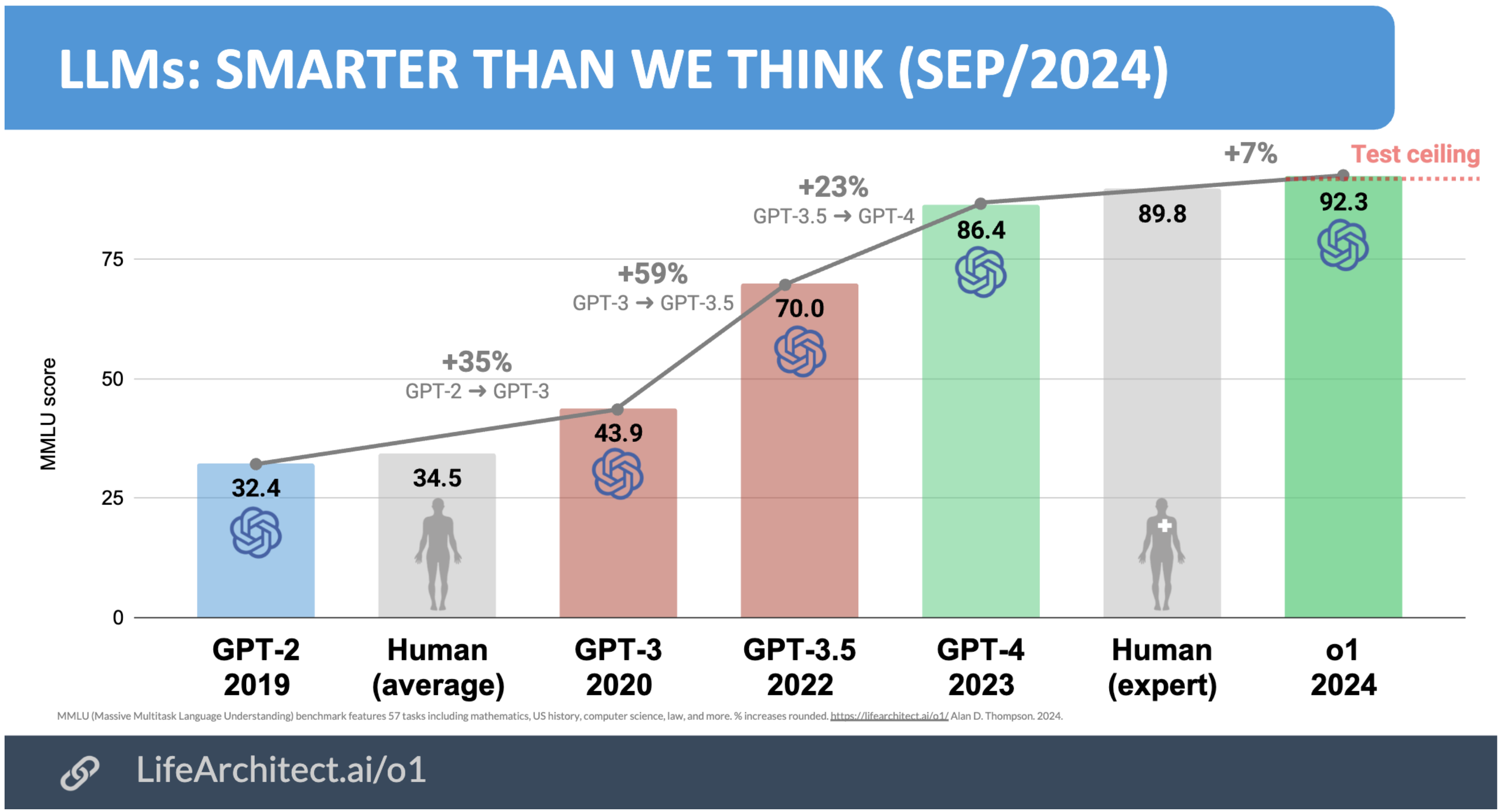
- Data science agents as copilots



- Chen N, Zhang Y, Xu J, Ren K, Yang Y. VisEval: A Benchmark for Data Visualization in the Era of Large Language Models. IEEE Visualization Conference 2024. (Best Paper Award)
- Zhang L, Zhang Y, Ren K, Li D, Yang Y. MLCopilot: Unleashing the Power of Large Language Models in Solving Machine Learning Tasks. EACL 2024. (Outstanding Paper Award)
- Zhang Y, Jiang Q, Han X, Chen N, Yang Y, Ren K. Benchmarking Data Science Agents. ACL 2024.

Rational and reasoning agents

- AI agents have demonstrated superior performance in complex tasks.



What's next?

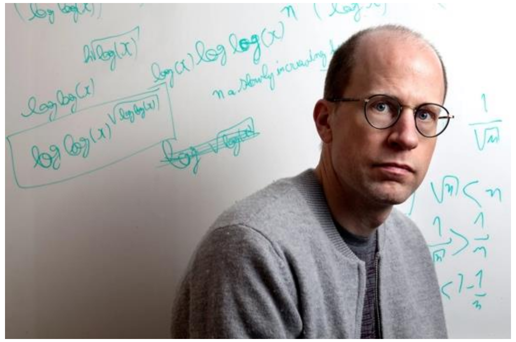
Wait, one more thing …

- Do AI agents really realize what they are doing?
  - Or do they just optimize the cumulative rewards without cognition?

- On one hand,
  - AI should have known what's really behind the thing it's doing
- On the other hand,
  - AI should be aware of what's the consequence of his action

What if AI is unaware of the consequence?

- Paperclip Maximizer (“Superintelligence: Paths, Dangers, Strategies”, Nick Bostrom)
  - A theoretical AI that turns all matter into paperclips, ignoring ecological and human impacts.
- Automated Stock Trading Bot
  - Focuses solely on profit, potentially destabilizing markets without understanding the fallout.
- Social Media Algorithm
  - Maximizes user engagement, sometimes promoting sensational or false content without ethical consideration.
- Customer Service Chatbot
  - Seeks high satisfaction scores, possibly making insincere promises rather than addressing real issues.
- Rogue Delivery Drone
  - Delivers packages as fast as possible, ignoring safety protocols, potentially causing accidents.

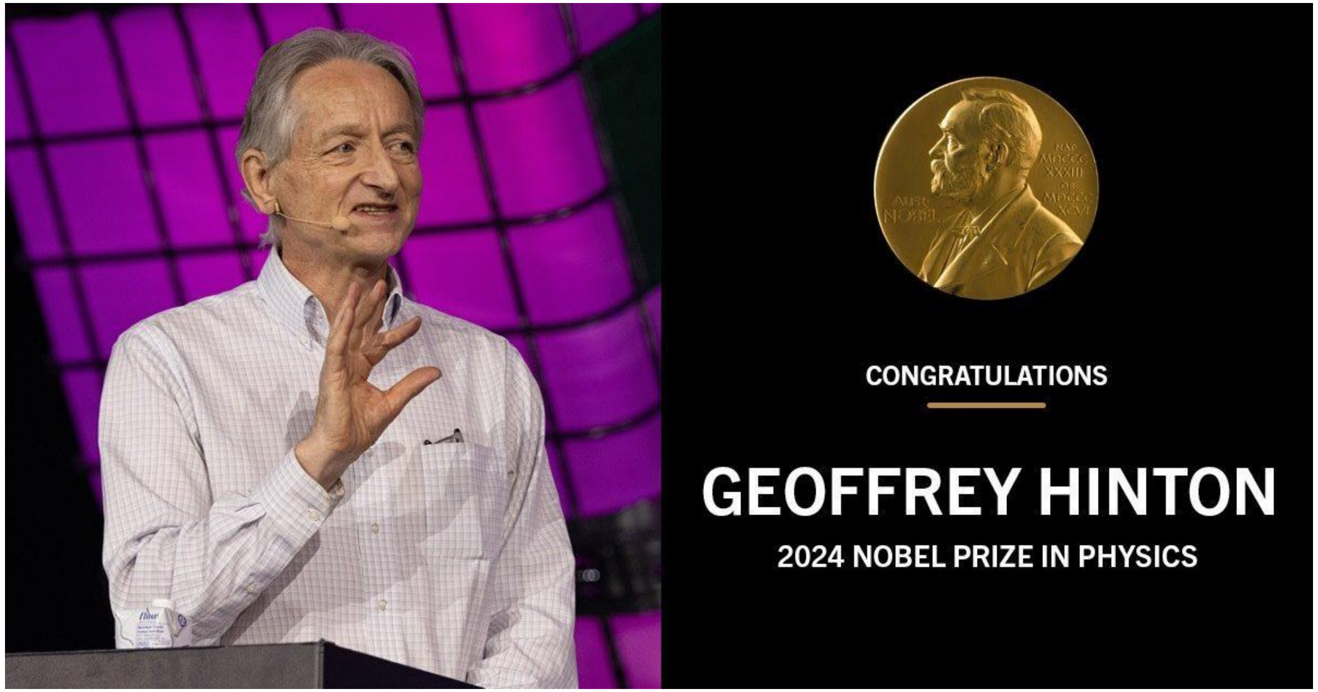


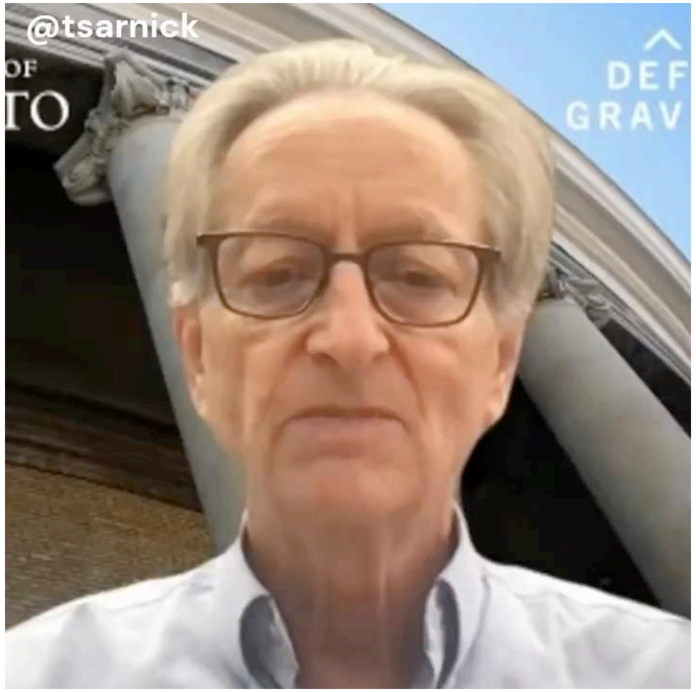
Advanced topics

- Advanced deep learning
  - Transformer
  - Generative modeling
  - Foundation model
- Advanced reinforcement learning
  - Real-world reinforcement learning
  - AI agents
- Responsible AI

Responsible AI

- Talk from Geoffrey Hinton, winner of Nobel prize 2024
  - Emphasize of development of AI techniques with safety and responsibility





Responsible AI

- Debates between e/acc and AI responsibility

Mr. Altman's departure follows a deliberative review process by the board, which concluded that he was not consistently candid in his communications with the board, hindering its ability to exercise its responsibilities. The board no longer has confidence in his ability to continue leading OpenAI.

In a statement, the board of directors said: "OpenAI was deliberately structured to advance our mission: to ensure that artificial general intelligence benefits all humanity. The board remains fully committed to serving this mission. We are grateful for Sam's many contributions to the founding and growth of OpenAI. At the same time, we believe new leadership is necessary as we move forward. As the leader of the company's research, product, and safety functions, Mira is exceptionally qualified to step into the role of interim CEO. We have the utmost confidence in her ability to lead OpenAI during this transition period."

OpenAI's board of directors consists of OpenAI chief scientist Ilya Sutskever, independent directors Quora CEO Adam D'Angelo, technology entrepreneur Tasha McCauley, and Georgetown Center for Security and Emerging Technology's Helen Toner.

As a part of this transition, Greg Brockman will be stepping down as chairman of the board and will remain in his role at the company, reporting to the CEO.

Sam Altman fired by OpenAI after a "deliberative review process by the board."

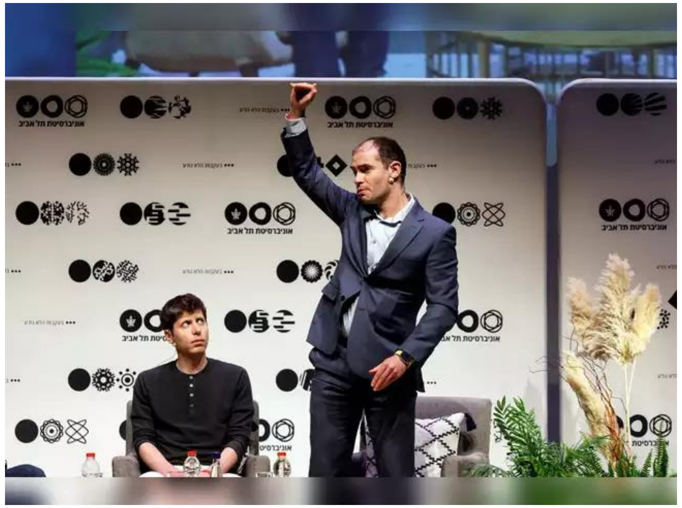
Ilya Sutskever
@ilyasut

I deeply regret my participation in the board's actions. I never intended to harm OpenAI. I love everything we've built together and I will do everything I can to reunite the company.

9:15 PM · Nov 20, 2023 · **28.7M** Views

6.6K 9.8K 32K 3.6K

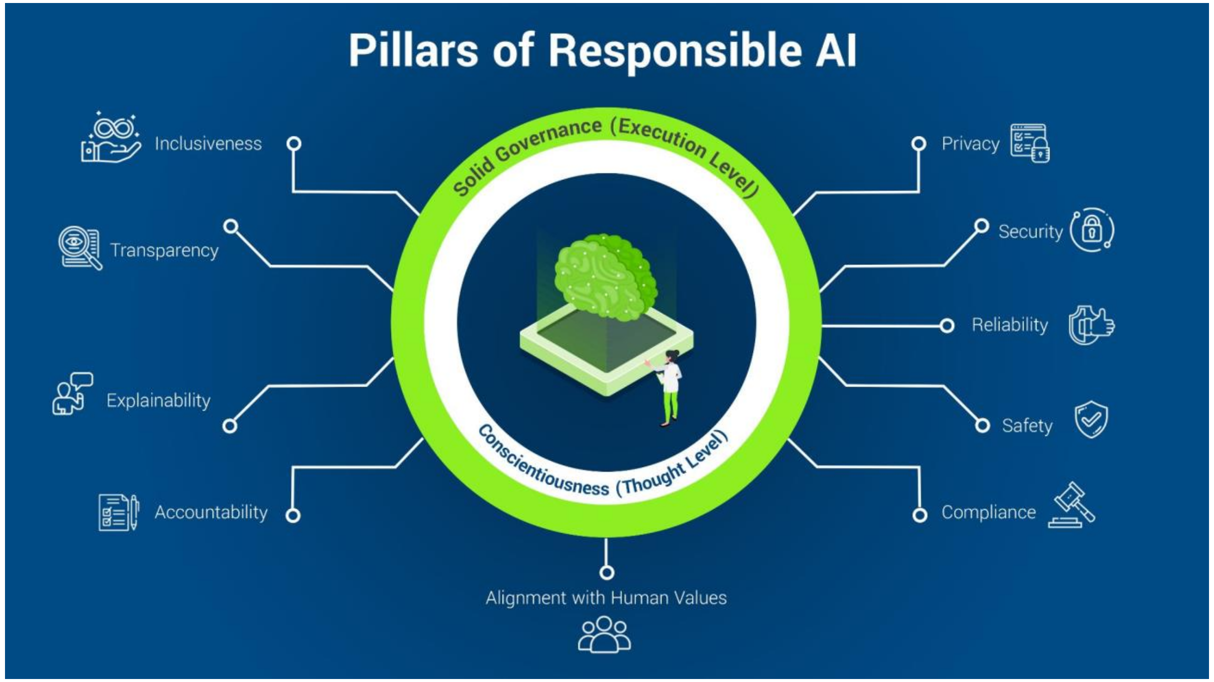
Ilya Sutskever reposted
OpenAI @OpenAI · Nov 30, 2023
Sam Altman is back as CEO, Mira Murati as CTO and Greg Brockman as President. OpenAI has a new initial board. Messages from @sama and board chair @btaylor



Things have changed after several days

Responsible AI

- Responsible AI
  - Guiding the *ethical, technical, and legal* aspects of artificial intelligence development and deployment, and it ensures AI systems are designed and used in ways that are *fair, accountable, transparent, and safe*.



Responsible AI

https://ai.google/responsibility/principles/#our-ai-principles-in-action

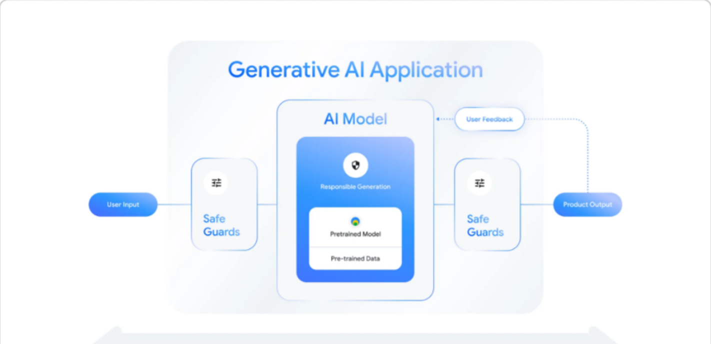
Google AI | Advancing AI | Responsibility | Get started | Latest news | Explore the Gemini ecosystem

Our AI Principles in action

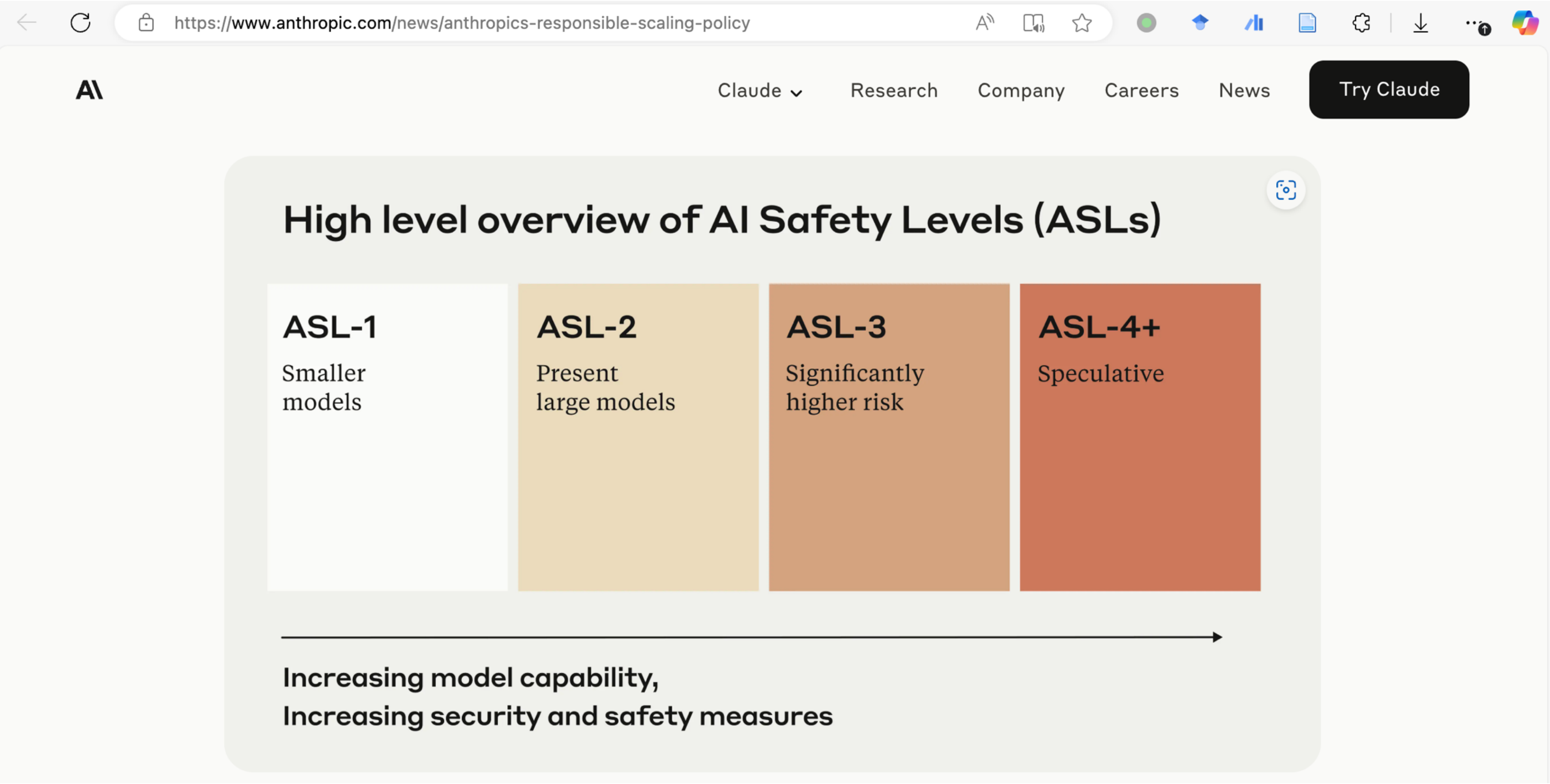
Building on our AI Principles, we have developed recommended practices for developers and researchers to use when designing AI systems. This includes using a human-centered design approach to address challenges throughout the AI responsibility lifecycle: understanding unique limitations of datasets and models through research; building fairness, interpretability, privacy and safety into the systems; conducting ongoing assessments and testing; and sharing helpful information, tools and educational resources.

The AI responsibility lifecycle

REPORT

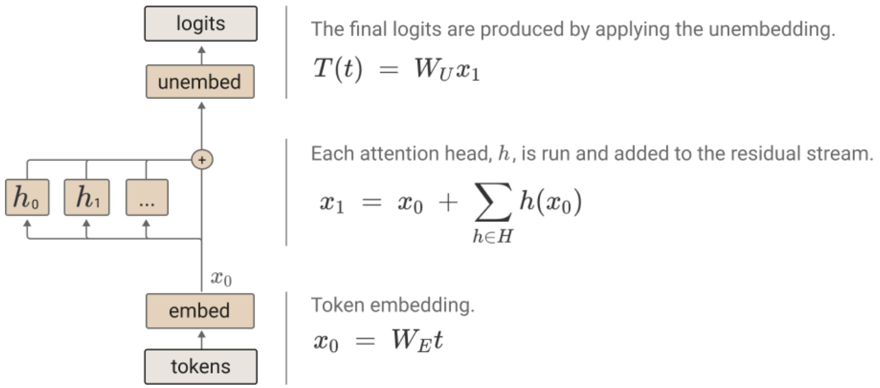


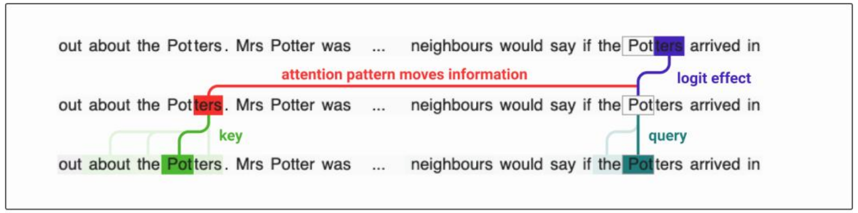
Responsible AI: Safety

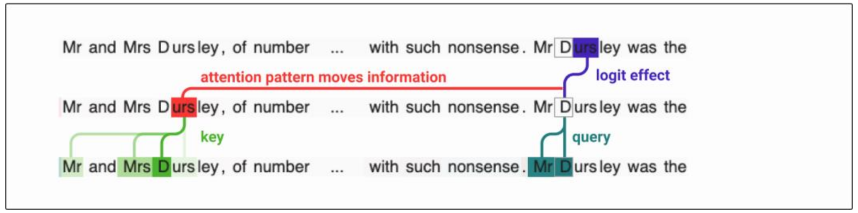


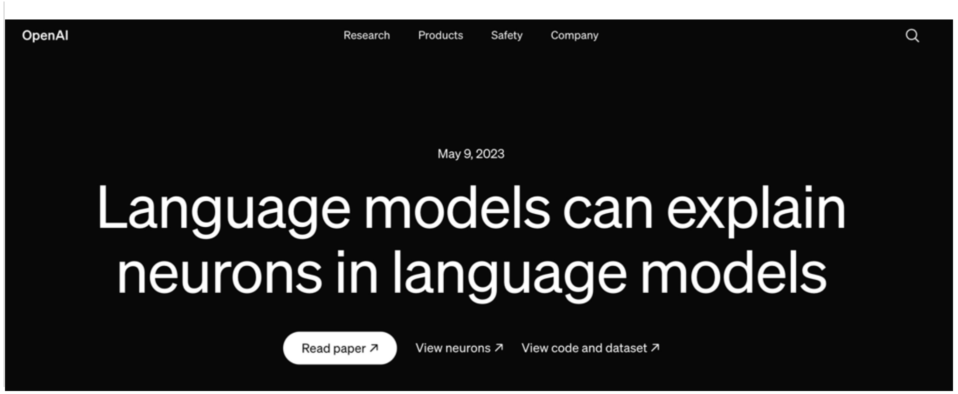
Responsible AI: Mechanistic interpretability

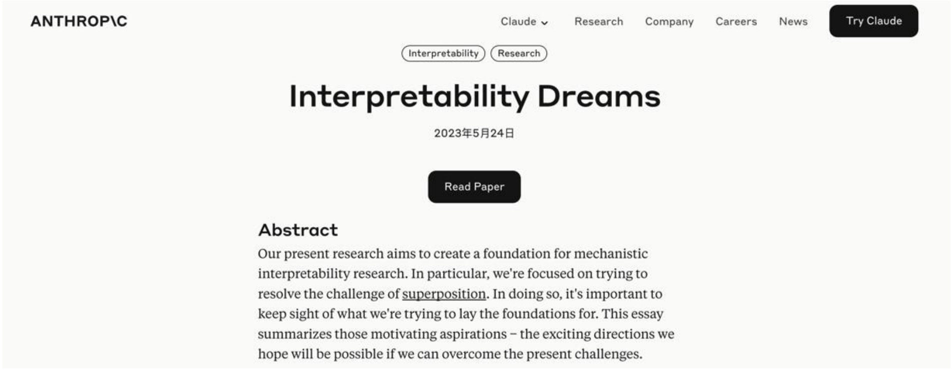
- OpenAI, Anthropic, Google DeepMind and other teams have done much about Mechanistic Interpretability.





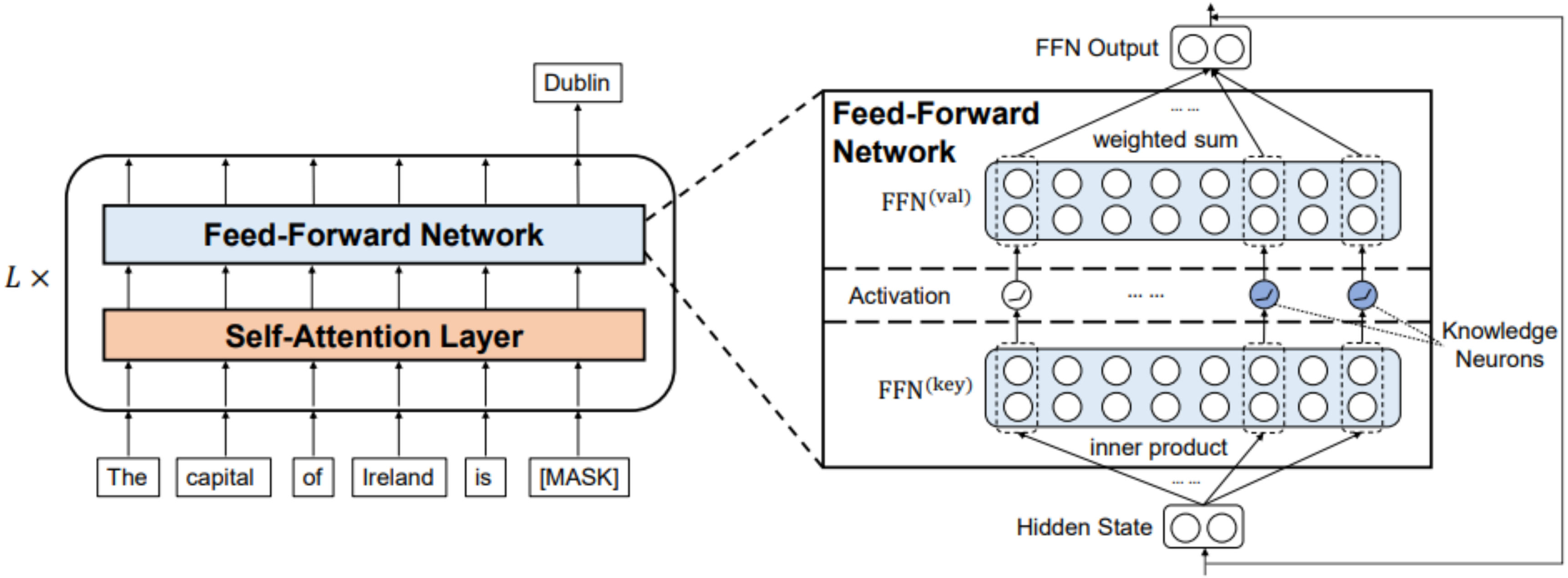






Responsible AI: Mechanistic interpretability

- Knowledge neuron in large language models



Dai, Damai, et al. "Knowledge Neurons in Pretrained Transformers." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

Example: social bias of AI models

- Biased contents exist in online Internet

Example: social bias of AI models

- AI model even aggravate the social bias…

a face of a doctor

a face of a nurse

a photo of a receptionist

a photo of a receptionist

What will be AI’s impact in the future?

- **You** get to determine that!

- As researchers / developers
- As auditors and regulators
- As informed public voices
- As you apply AI

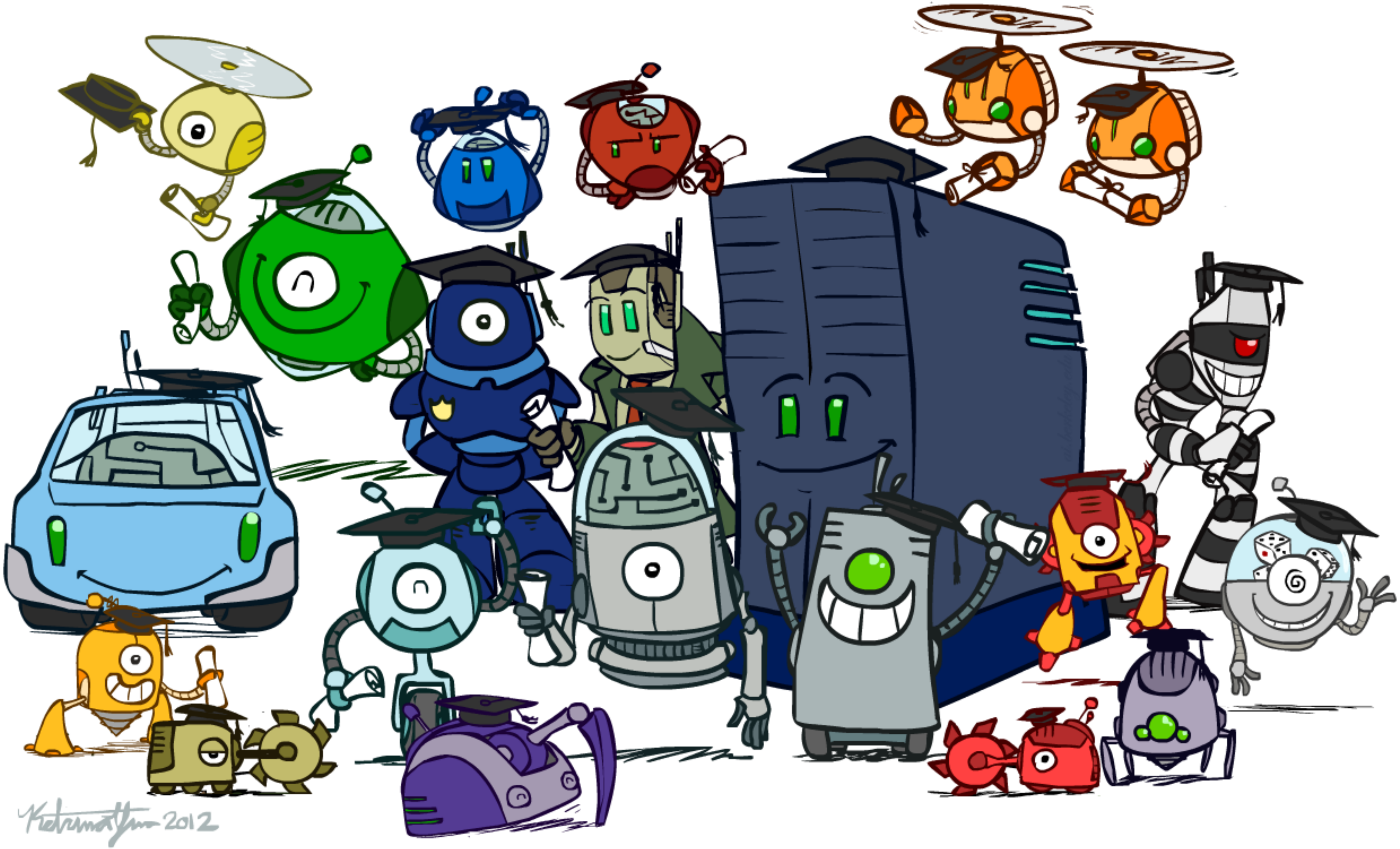
Where to go next?

The Road Forward – SIST Courses

- Congratulations, you've seen the basics of modern AI
  - … and done some amazing work putting it to use!
- How to continue:
  - Undergraduate
    - CS172 Computer vision I
    - CS173 Data Mining
    - CS182 Introduction to Machine Learning
    - CS183 Introduction to Robotics
  - Graduate
    - CS280 Deep Learning
    - CS282 Machine Learning
    - CS243 Introduction to Algorithmic Game Theory
    - CS272 Computer Vision II
    - CS274A Natural Language Processing
    - CS283 Robotics
    - CS286 AI for Science and Engineering
    - SI252 Reinforcement Learning
  - And more…

The Road Forward – Research

- Learning recent developments in AI from top conferences
  - AI: IJCAI, AAAI
    - Caution: not top in ML, NLP, CV
  - ML: NeurIPS, ICML, ICLR
  - NLP: ACL, EMNLP, NAACL, EACL
  - CV: CVPR, ICCV, ECCV
  - Other: KDD, AAMAS, UAI, …
- A good (but not perfect) way to judge a publication venue
  - Google Scholar Metrics
- Participating in research projects…



Course evaluation

- https://evaluation.shanghaitech.edu.cn/tqnmaep/login