

Yonsei University Applied Statistics Invited Seminar

LLM & genAI - Technology, Industry, Market, and Some Important Questions

Sunghee Yun

Co-founder - AI Technology & Product Strategy

Erudio Bio, Inc.

About Speaker

- *Co-founder - AI Technology & Product Strategy @ Erudio Bio, CA, USA*
- Advisory Professor, Electrical Engineering and Computer Science @ DGIST
- Adjunct Professor, Electronic Engineering Department @ Sogang University
- Technology Consultant @ Gerson Lehrman Group (GLG), NYC, USA
- *Co-founder / CTO @ Gauss Labs, Palo Alto, CA, USA – 2023*
- Senior Applied Scientist @ Amazon, Vancouver, Canada – 2020
- Principal Engineer @ Software R&D Center of DS Division - Samsung – 2017
- Principal Engineer @ Strategic Marketing Team of Memory Business Unit – 2016
- Principal Engineer @ DT Team of DRAM Development Lab. - Samsung – 2015
- Senior Engineer @ CAE Team - Samsung – 2012
- M.S. & Ph.D. - Electrical Engineering @ Stanford University – 2004
- B.S. - Electrical Engineering @ Seoul National University – 1998

Highlight of career journey

- B.S. in EE @ SNU, M.S. & Ph.D. in EE @ Stanford Univ.
 - *Convex Optimization - theory / algorithms / applications - under supervision of Prof. Stephen P. Boyd*
- Principal Engineer @ Memory Design Technology Team
 - AI & optimization partnering with *DRAM/NAND Design/Process/Test Teams*
- Senior Applied Scientist @ Amazon
 - *S-Team Goal (Bezos's) project - better customer shopping experience via Amazon shopping app using AI - increased sales by \$200M*
- Co-founder / CTO & Chief Applied Scientist @ Gauss Labs
 - *lead develop & productionize industrial AI products & technology roadmapping, market/product/investment strategies*
- Co-founder - AI Technology & Product Strategy @ Erudio Bio
 - *biotech - AI technology & product strategy*

Today

- AI trend and technology
 - large language model (LLM)
 - *attention turns out to be way more efficient*
. . . than even original authors envisioned!
 - genAI & multimodal learning - models & applications
- industry and business impacts
 - business applications & products
 - AI research, AI market, investment on AIs in Silicon Valley
- some important topics & questions around AI
 - why DL works amazingly well?
 - AI ethics, law, biases, consciousness
 - utopia / dystopia and many others

Takeaways and questions

- purpose of this talk is answer questions such as . . .
 - what is *secret sauce of LLM*?
 - why is multimodal learning promising technology?
 - AI trend in industry & academia
 - *industry, market & social impacts of AI*
- and *make audience curious about topics such as . . .*
 - what are things that we should be cautious of about AIs?
 - how can we prevent potential harms of AI?
 - how can / should we prepare for known & unknown changes brought by AI?
 - questions like . . . is AI intelligent? knowledgeable? has it consciousness?

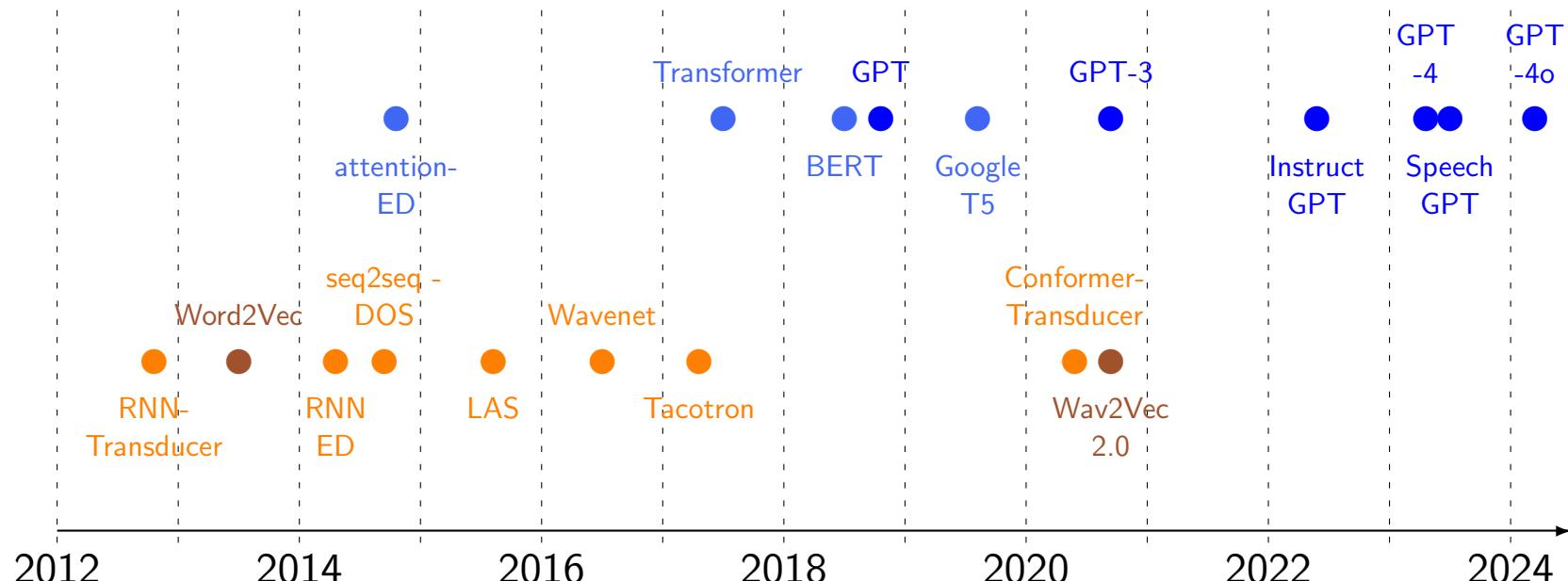
Technology

Language Models

History of language models

- bag of words - first introduced – 1954
- word embedding – 1980
- RNN based models - conceptualized by David Rumelhart – 1986
- LSTM (based on RNN) – 1997
- 380M-sized seq2seq model using LSTMs proposed – 2014
- 130M-sized seq2seq model using gated recurrent units (GRUs) – 2014
- Transformer - Attention is All You Need - A. Vaswani et al. @ Google – 2017
 - 100M-sized encoder-decoder multi-head attention model for machine translation
 - non-recurrent architecture, handle arbitrarily long dependencies
 - parallelizable, *simple* (linear-mapping-based) attention model

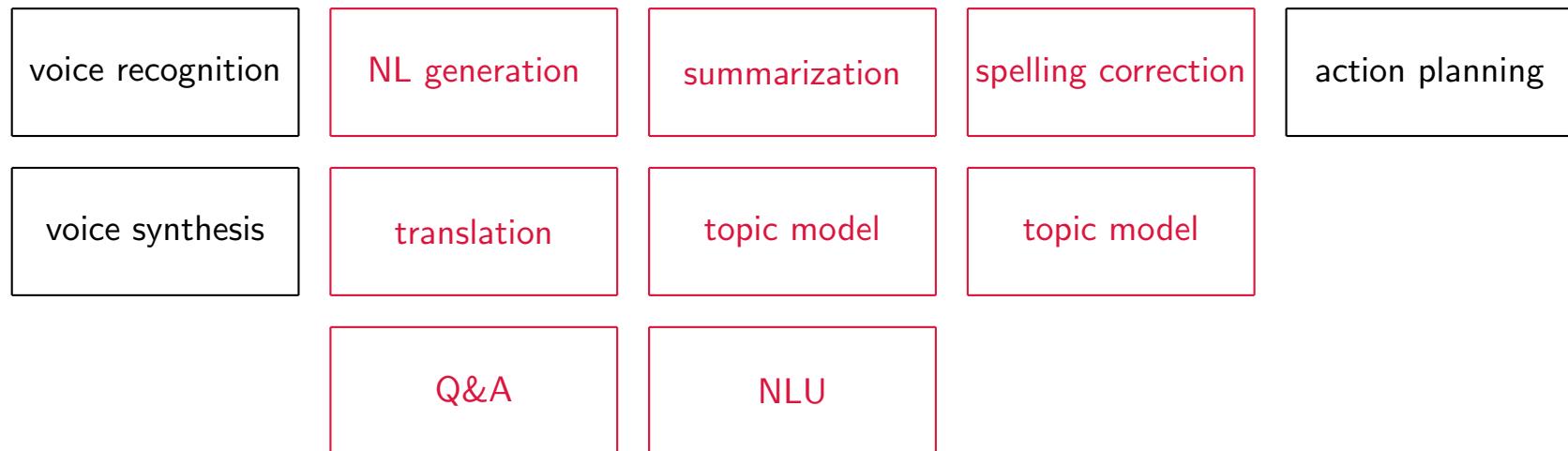
Recent advances in speech & language processing



- LAS: listen, attend, and spell, ED: encoder-decoder, DOS: decoder-only structure

Types of language models

- many of language models have **common requirements** - language representation learning
- can be learned via pre-training *high performing model* and fine-tuning/transfer learning/domain adaptation
- this *high performing model* learning essential language representation *is* (language) foundation model
 - actually, same for other types of learning, e.g., CV



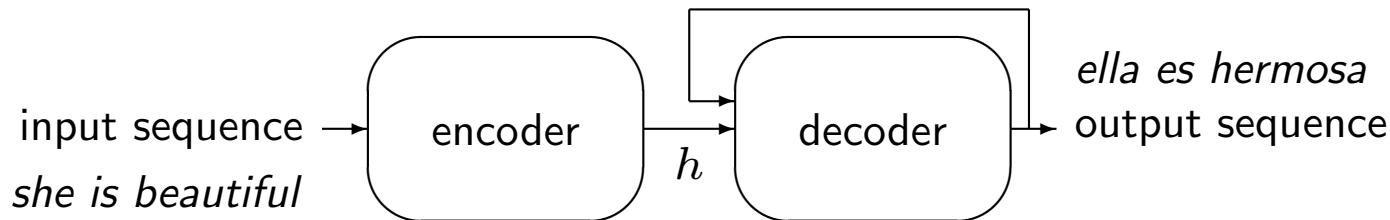
NLP market size

- global NLP market size estimated at USD 16.08B in 2022, is expected to hit USD 413.11B by 2032 - *CAGR of 38.4%*
- in 2022
 - north america NLP market size valued at USD 8.2B
 - high tech and telecom segment accounted revenue share of over 23.1%
 - healthcare segment held a 10% market share
 - (by component) solution segment hit 76% revenue share
 - (deployment mode) on-premise segment generated 56% revenue share
 - (organizational size) large-scale segment contributed highest market share
- source - Precedence Research



RNN-type sequence to sequence (seq2seq) model

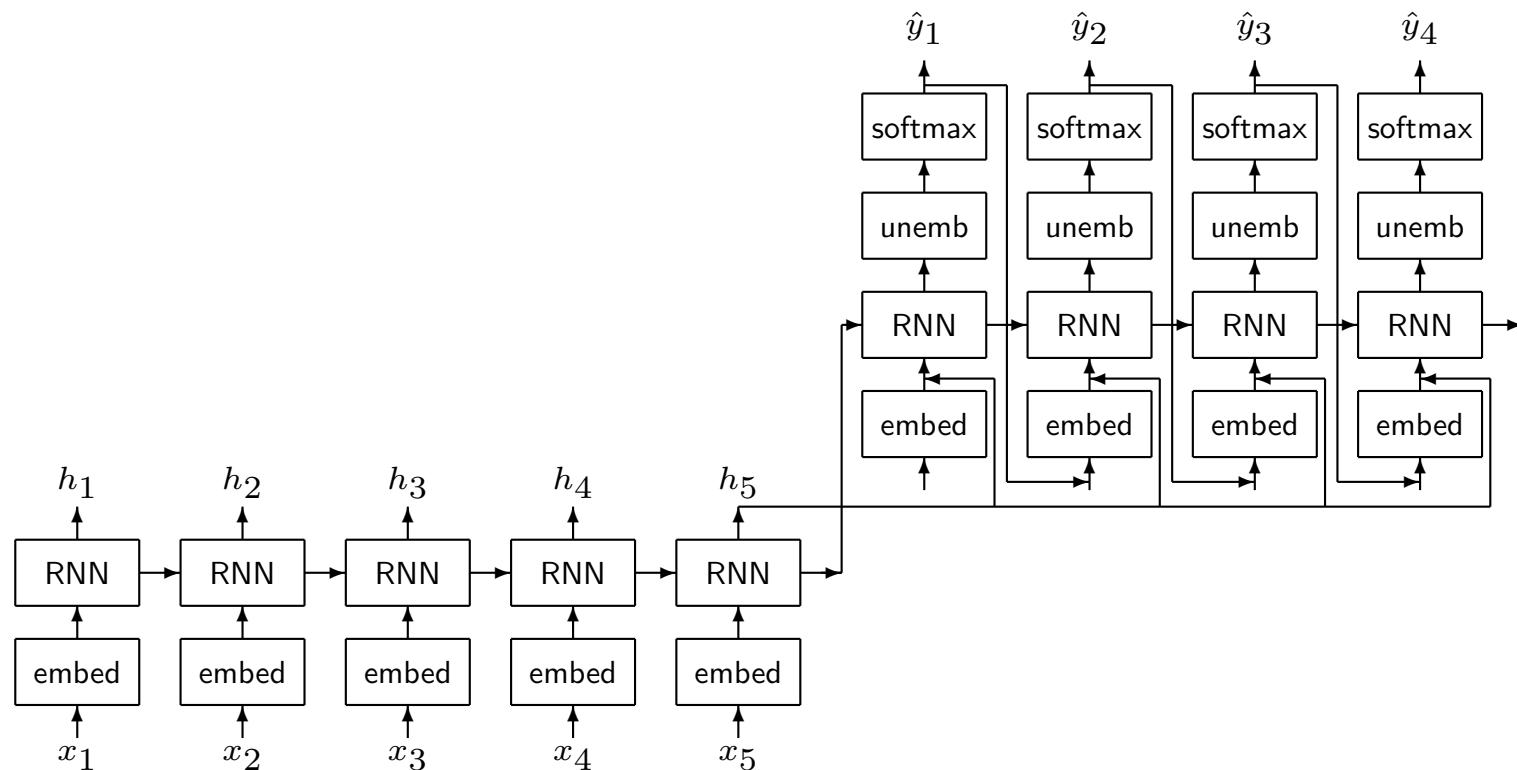
- seq2seq - take sequences as inputs and spit out sequences
- encoder-decoder architecture



- encoder & decoder is RNN-type model
- $h \in \mathbf{R}^n$ - hidden state - *fixed length* vector
- (try to) condense and store information of input sequence (losslessly) in (fixed-length) hidden states
 - finite hidden state - not flexible enough, *i.e.*, cannot handle arbitrarily large information
 - memory loss for long sequences
 - LSTM was promising fix, but with (inevitable) limits

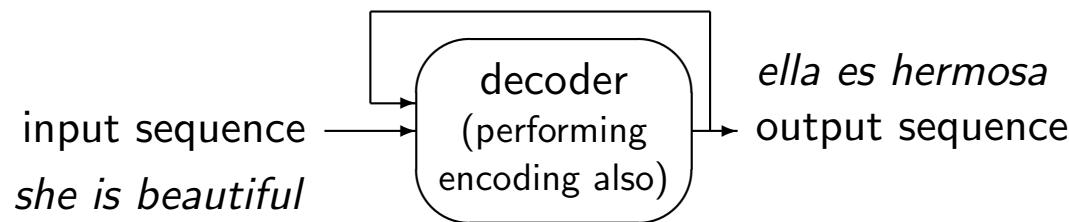
RNN-type encoder-decoder example

- RNN can be basic RNN, LSTM, GRU, etc.



Shared encoder/decoder model

- may use single structure to perform both encoding & decoding
- LLMs are built in this way



Large Language Models

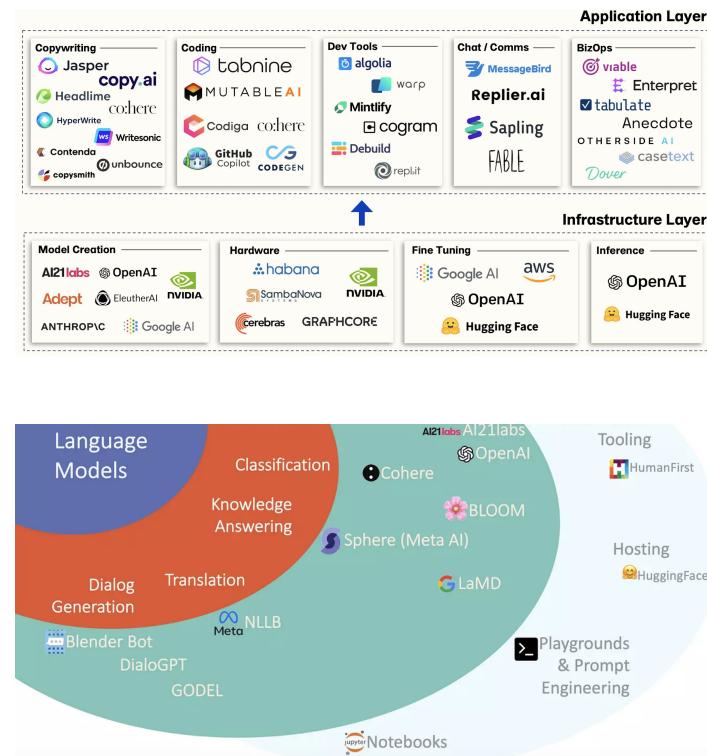
Large language model (LLM)

- LLM
 - type of AI aimed for NLP trained on massive corpus of texts & programming code
 - allow learn statistical relationships between words & phrases, *i.e.*, conditional probabilities
 - *amazing performance shocked everyone - unreasonable effectiveness of data (Halevy et al., 2009)*
 - applications
 - conversational AI agent / virtual assistant
 - machine translation / text summarization / content creation / sentiment analysis
 - code generation
 - market research / legal service / insurance policy / triange hiring candidates
 - + virtually infinite # of applications



LLMs

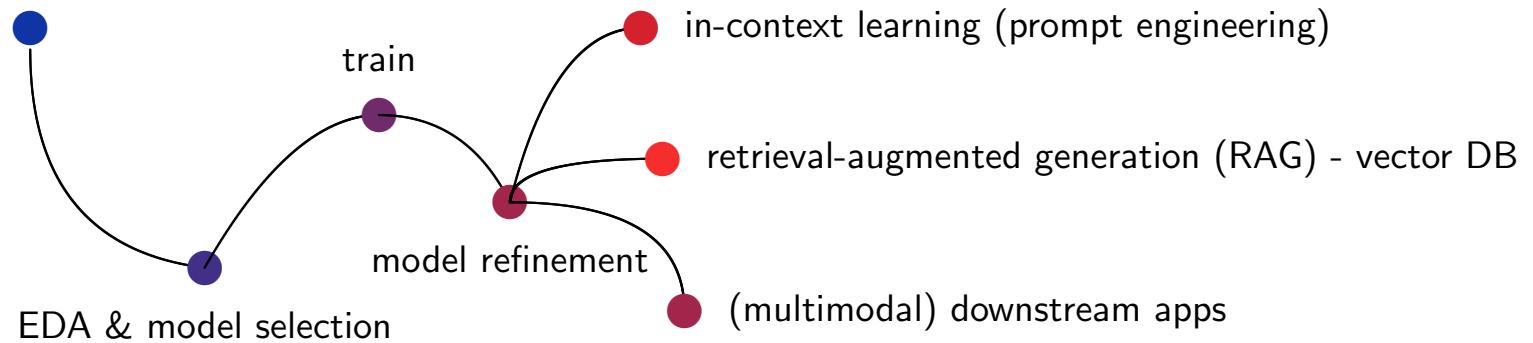
- Foundation Models
 - GPT-x/Chat-GPT - OpenAI, Llama-x - Meta, PaLM-x (Bard) - Google
- # parameters
 - generative pre-trained transformer (GPT) - GPT-1: 117M, GPT-2: 1.5B, GPT-3: 175B, GPT-4: 100T, GPT-4o: 200B
 - large language model Meta AI (Llama) - Llama1: 65B, Llama2: 70B, Llama3: 70B
 - scaling language modeling with pathways (PaLM) - 540B
- burns lots of cash on GPUs!
- applicable to many NLP & genAI applications



LLM building blocks

- data - trained on massive datasets of text & code
 - quality & size critical on performance
- architecture - GPT/Llama/Mistral
 - can make huge difference
- training - self-supervised/supervised learning
- inference - generates outputs
 - in-context learning, prompt engineering

goal and scope of LLM project



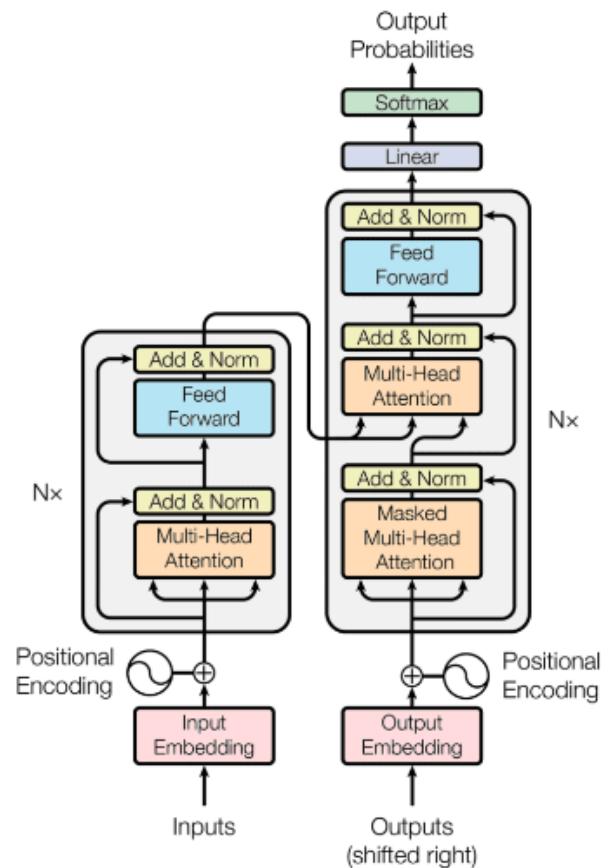
LLM architectural secret (or known) sauce

Transformer - simple parallelizable attention mechanism

A. Vaswani, et al. Attention is All You Need, 2017

Transformer architecture

- encoding-decoding architecture
 - input embedding space → multi-head & mult-layer representation space → output embedding space
- additive positional encoding - information regarding order of words @ input embedding
- multi-layer and multi-head attention followed by addition / normalization & feed forward (FF) layers
- *(relatively simple) attentions*
 - single-head (scaled dot-product) / multi-head attention
 - self attention / encoder-decoder attention
 - masked attention
- benefits
 - *evaluate dependencies between arbitrarily distant words*
 - has recurrent nature w/o recurrent architecture → parallelizable → fast w/ additional cost in computation



Single-head scaled dot-product attention

- values/keys/queries denote value/key/query *vectors*, d_k & d_v are lengths of keys/queries & vectors
- we use *standard* notions for matrices and vectors - not transposed version that (almost) all ML scientists (wrongly) use
- output: weighted-average of values where weights are attentions among tokens
- assume n queries and m key-value pairs

$$Q \in \mathbf{R}^{d_k \times n}, K \in \mathbf{R}^{d_k \times m}, V \in \mathbf{R}^{d_v \times m}$$

- attention! outputs n values (since we have n queries)

$$\text{Attention}(Q, K, V) = V \text{softmax} \left(K^T Q / \sqrt{d_k} \right) \in \mathbf{R}^{d_v \times n}$$

- *much simpler attention mechanism than previous work*
 - attention weights were output of complicated non-linear NN

Single-head - close look at equations

- focus on i th query, $q_i \in \mathbf{R}^{d_k}$, $Q = [\quad - \quad q_i \quad - \quad] \in \mathbf{R}^{d_k \times n}$
- assume m keys and m values, $k_1, \dots, k_m \in \mathbf{R}^{d_k}$ & $v_1, \dots, v_m \in \mathbf{R}^{d_v}$

$$K = [\quad k_1 \quad \cdots \quad k_m \quad] \in \mathbf{R}^{d_k \times m}, V = [\quad v_1 \quad \cdots \quad v_m \quad] \in \mathbf{R}^{d_v \times m}$$

- then

$$K^T Q / \sqrt{d_k} = \begin{bmatrix} & & \vdots \\ - & k_j^T q_i / \sqrt{d_k} & - \\ & & \vdots \end{bmatrix}$$

e.g., dependency between i th output token and j th input token is

$$a_{ij} = \exp \left(k_j^T q_i / \sqrt{d_k} \right) / \sum_{j=1}^m \exp \left(k_j^T q_i / \sqrt{d_k} \right)$$

- value obtained by i th query, q_i in $\text{Attention}(Q, K, V)$

$$a_{i,1}v_1 + \cdots + a_{i,m}v_m$$

Multi-head attention

- evaluate h single-head attentions (in parallel)
- d_e : dimension for embeddings
- embeddings

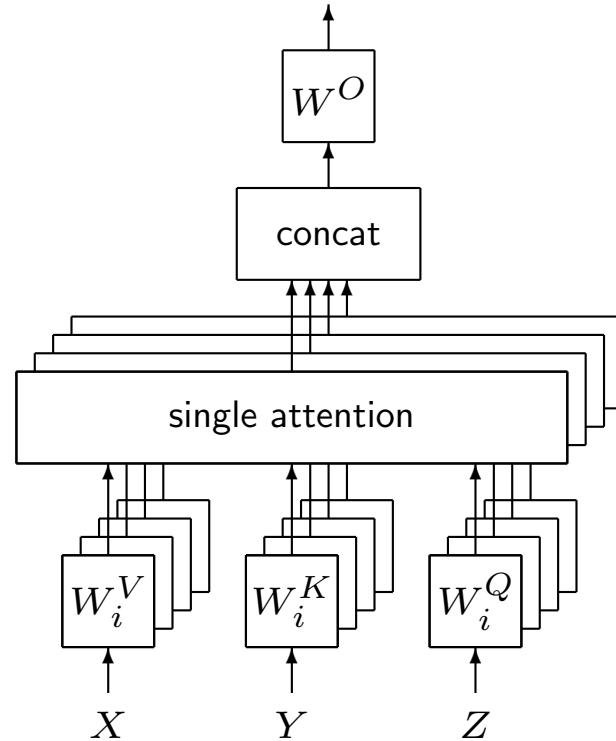
$$X \in \mathbb{R}^{d_e \times m}, Y \in \mathbb{R}^{d_e \times m}, Z \in \mathbb{R}^{d_e \times n}$$

e.g., n : input sequence length & m : output sequence length in machine translation

- h key/query/value weight matrices: $W_i^K, W_i^Q \in \mathbb{R}^{d_k \times d_e}$, $W_i^V \in \mathbb{R}^{d_v \times d_e}$ ($i = 1, \dots, h$)
- linear output layers: $W^O \in \mathbb{R}^{d_e \times hdv}$
- *multi-head attention!*

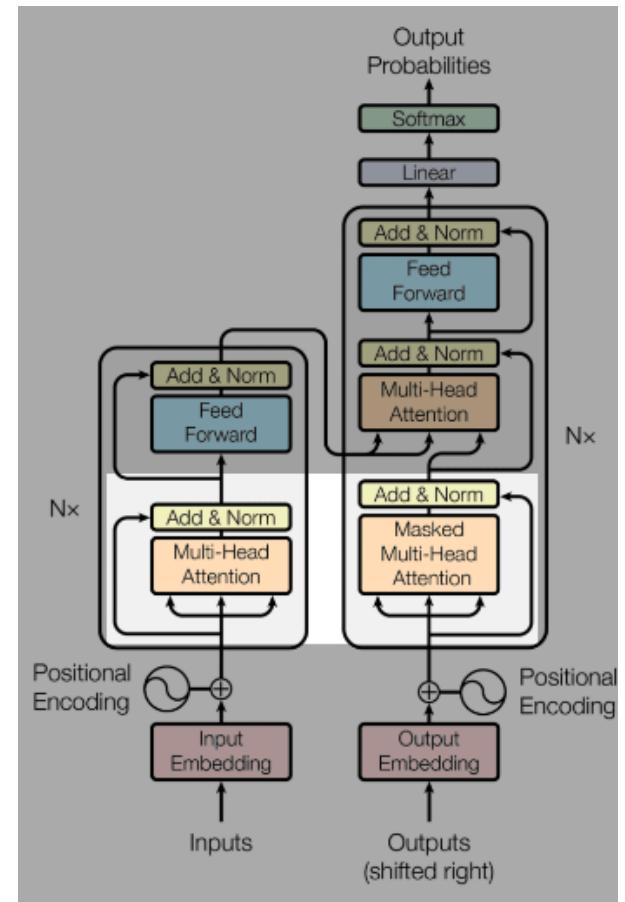
$$W^O \begin{bmatrix} A_1 \\ \vdots \\ A_h \end{bmatrix} \in \mathbb{R}^{d_e \times n},$$

$$A_i = \text{Attention}(W_i^Q Z, W_i^K Y, W_i^V X) \in \mathbb{R}^{d_v \times n}$$



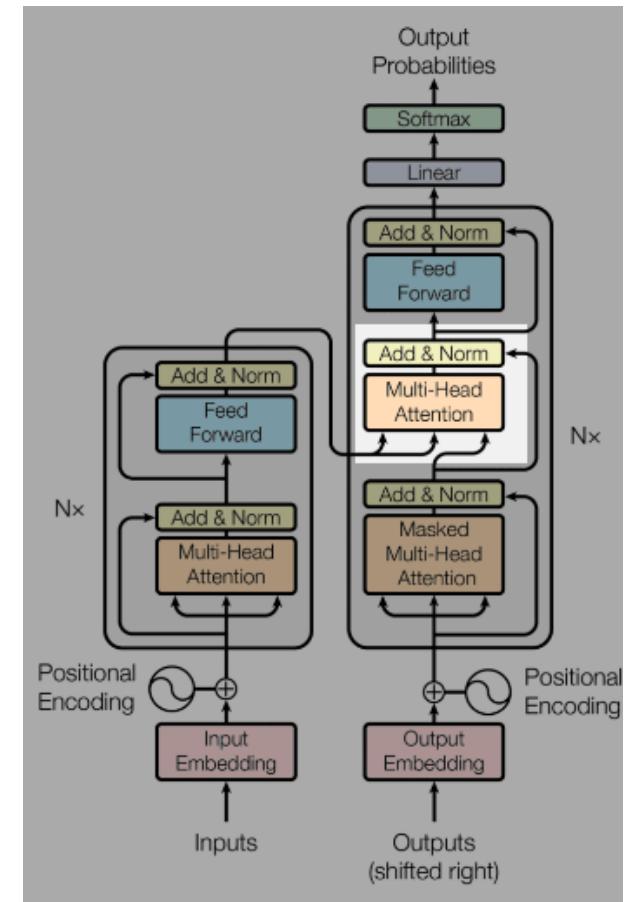
Self attention

- $m = n$
- encoder
 - keys & values & queries (K, V, Q) come from same place (from previous layer)
 - every token attends to every other token in input sequence
- decoder
 - keys & values & queries (K, V, Q) come from same place (from previous layer)
 - every token attends to other tokens up to that position
 - prevent leftward information flow to right to preserve causality
 - assign $-\infty$ for illegal connections in softmax (masking)



Encoder-decoder attention

- m : length of input sequence
- n : length of output sequence
- n queries (Q) come from previous decoder layer
- m keys / m values (K, V) come from output of encoder
- every token in output sequence attends to every token in input sequence

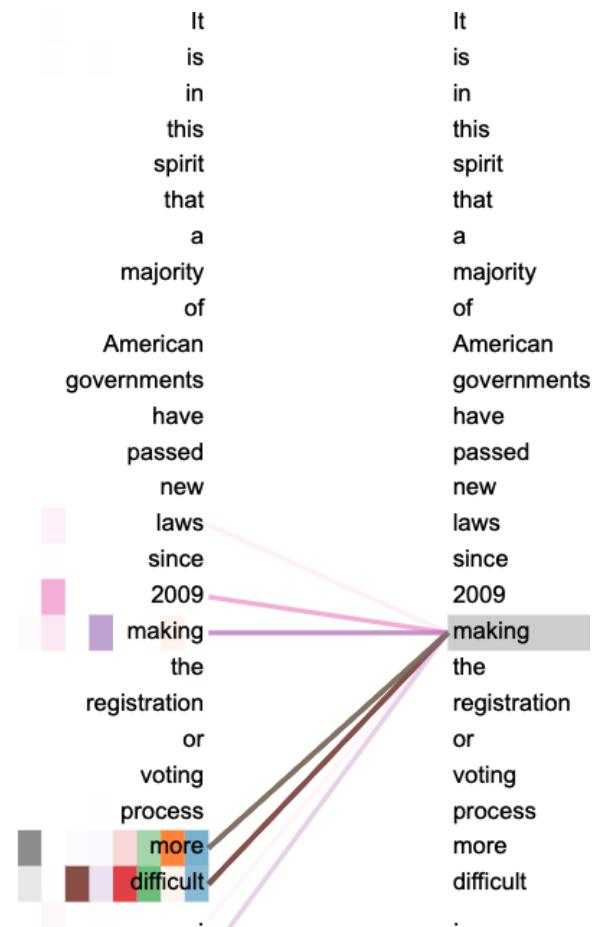


Visualization of self attentions

example sentence

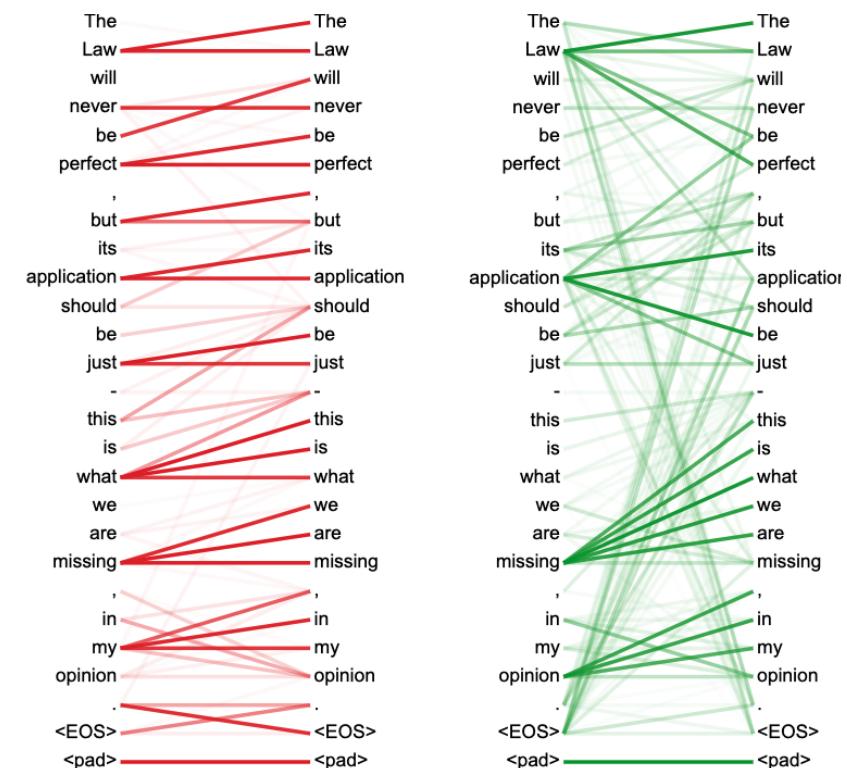
"It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult."

- self attention of encoder (of a layer)
 - right figure
 - show dependencies between "making" and other words
 - different columns of colors represent different heads
 - "making" has strong dependency to "2009", "more", and "difficult"

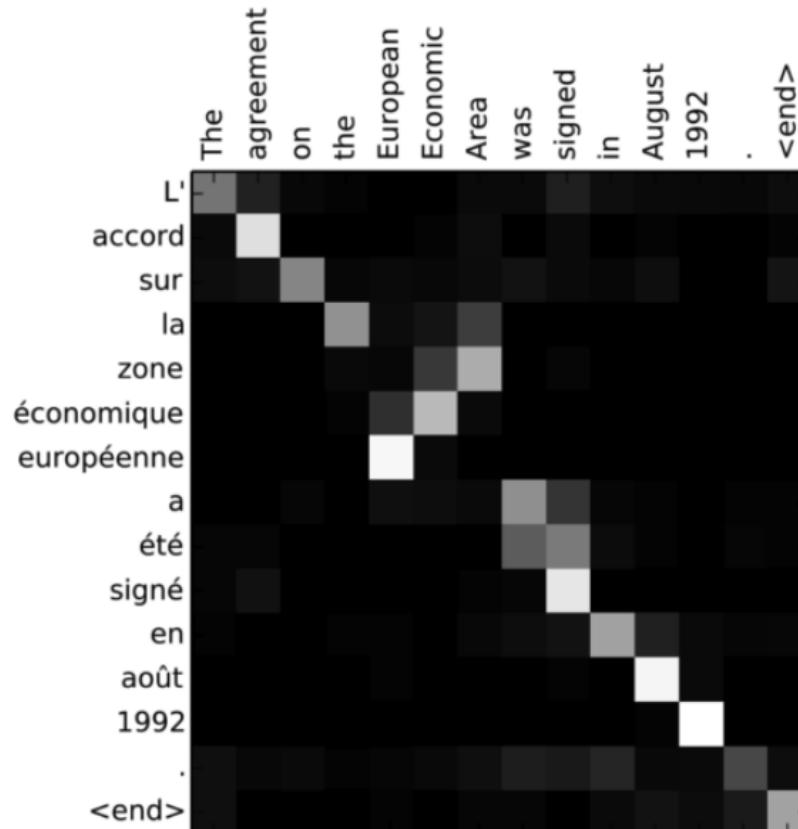


Visualization of multi-head self attentions

- self attentions of encoder for two heads (of a layer)
 - different heads represent different structures
→ advantages of multiple heads
 - multiple heads work together to collectively yield good results
 - dependencies *not* have absolute meanings (like embeddings in collaborative filtering)
 - randomness in resulting dependencies exists due to stochastic nature of ML training



Visualization of encoder-decoder attentions



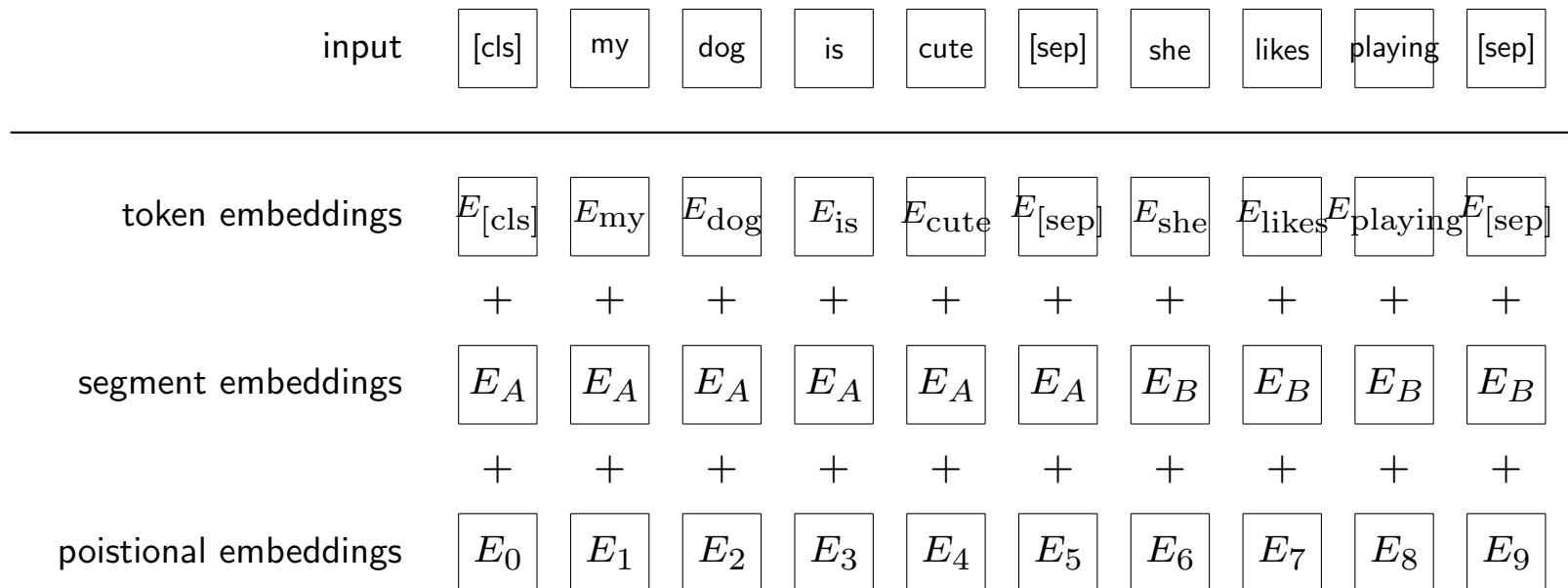
- machine translation: English → French
 - input sentence: “The agreement on the European Economic Area was signed in August 1992.”
 - output sentence: “L’ accord sur la zone économique européenne a été signé en août 1992.”
- encoder-decoder attention reveals relevance between
 - European ↔ européenne
 - Economic ↔ européenne
 - Area ↔ zone

Model complexity

- computational complexity
 - n : sequence length, d : embedding dimension
 - complexity per layer - self-attention: $\mathcal{O}(n^2d)$, recurrent: $\mathcal{O}(1)$
 - sequential operations - self-attention: $\mathcal{O}(1)$, recurrent: $\mathcal{O}(n)$
 - maximum path length - self-attention: $\mathcal{O}(1)$, recurrent: $\mathcal{O}(n)$
- *massive parallel processing, long context windows*
 - makes NVidia more competitive, hence profitable!
 - makes SK Hynix prevail HBM market!

Derivatives of Transformer - BERT

- Bidirectional Encoder Representations from Transformers [Devlin et al., 2019]
- pre-train deep bidirectional representations from unlabeled text
- fine-tunable for multiple purposes



Multimodal learning

- understand information from multiple modalities, *e.g.*, text, images, audio, and video
- representation learning
 - language representation + image / video / text / audio representation
 - learn multimodal representations together
- outputs
 - captions for images, videos with narration, musics with lyrics
- collaboration among different modalities
 - understand image world (open system) using language (closed system)



Implications of success of LLMs

- (very) many researchers change gears towards LLM
 - from computer vision (CV), speech, music, video, even reinforcement learning
- *LLM is not (only) about languages . . .*
 - humans have . . .
 - evolved and optimized (natural) language structures for eons
 - handed down knowledge using natural languages for thousands of years
 - natural language optimized (in human brains) through *thousands of generations by evolution*
 - *can connect non-linguistic world (open system) using language structures (closed system)*

Challenges in LLMs

- *hallucination - can give entirely plausible outcome that is false*
- data poison attack
- unethical or illegal content generation
- huge resource necessary for both training & inference
- model size - need compact models
- outdated knowledge - can be couple of years old
- lack of reproducibility
- *biases - more on this later . . .*

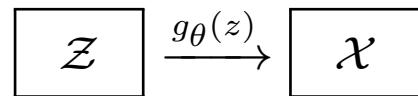
do not, though, focus on downsides but on *infinite possibilities!*

- it evolves like internet / mobile / electricity
- only “tip of the iceberg” found & revealed

Generative AI

Generative AI (genAI)

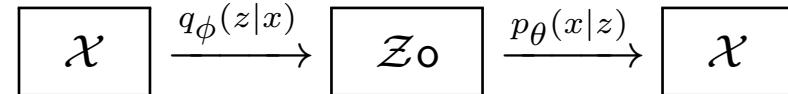
- definition of generative model



- *generate samples in original space, \mathcal{X} , from samples in latent space, \mathcal{Z}*
- g_θ is parameterized model *e.g.*, CNN / RNN / Transformer / diffuction-based model
- training
 - finding θ that minimizes/maximizes some (statistical) loss/merit function so that $\{g_\theta(z)\}_{z \in \mathcal{Z}}$ generates plausible point in \mathcal{X}
- inference
 - random samples z to generated target samples $x = g_\theta(z)$
 - *e.g.*, image, text, voice, music, video

VAE - early genAI model

- variational auto-encoder (VAE) [Kingma and Welling, 2019]



- log-likelihood & ELBO - for any $q_\phi(z|x)$

$$\begin{aligned}
 \log p_\theta(x) &= \mathbf{E}_{z \sim q_\phi(z|x)} \log p_\theta(x) = \mathbf{E}_{z \sim q_\phi(z|x)} \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \cdot \frac{q_\phi(z|x)}{p_\theta(z|x)} \\
 &= \mathcal{L}(\theta, \phi; x) + D_{KL}(q_\phi(z|x) \| p_\theta(z|x)) \geq \mathcal{L}(\theta, \phi; x)
 \end{aligned}$$

- (indirectly) maximize likelihood by maximizing evidence lower bound (ELBO)

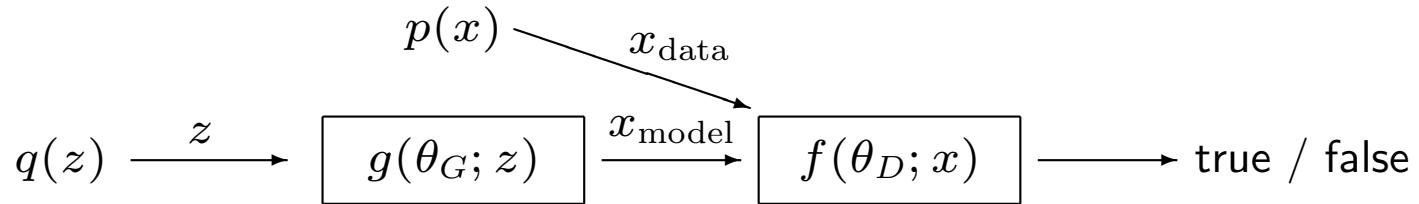
$$\mathcal{L}(\theta, \phi; x) = \mathbf{E}_{z \sim q_\phi(z|x)} \log \frac{p_\theta(x, z)}{q_\phi(z|x)}$$

- generative model

$$p_\theta(x|z)$$

GAN - early genAI model

- generative adversarial networks (GAN) [Goodfellow et al., 2014]



- value function

$$V(\theta_D, \theta_G) = \mathbf{E}_{x \sim p(x)} \log f(\theta_D; x)) + \mathbf{E}_{z \sim q(z)} \log(1 - f(\theta_D; g(\theta_G; z)))$$

- modeling via playing min-max game

$$\min_{\theta_G} \max_{\theta_D} V(\theta_D, \theta_G)$$

- generative model

$$g(\theta_G; z)$$

- variants: conditional / cycle / style / Wasserstein GAN

genAI - LLM

- *maximize conditional probability*

$$\underset{\theta}{\text{maximize}} \ d(p_{\theta}(x_t|x_{t-1}, x_{t-2}, \dots), p_{\text{data}}(x_t|x_{t-1}, x_{t-2}, \dots))$$

where $d(\cdot, \cdot)$ distance measure between probability distributions

- previous sequence: x_{t-1}, x_{t-2}, \dots
- next token: x_t
- p_{θ} represented by (extremely) complicated model
 - e.g., containing multi-head & multi-layer Transformer architecture inside
- model parameters, e.g., for Llama2

$$\theta \in \mathbf{R}^{70,000,000,000}$$

AI Applications

genAI applications

- ChatGPT, Cohere
- Anthropic, Dolly, Mosaic MPT
- Stable Diffusion
- Midjourney, DALL-E, LLaMA 2
- Mistral AI, Amazon Bedrock, and Falcon



ChatGPT & VR/AR

- new appropriately to teaching
- power of ChatGPT and VR/AR unlocks immersive learning
 - *learning language* - immersive VR environment provides immediate feedback, responding to inquiries & interactive discussions
 - *medical education* - experience diagnosing and treating patients in lifelike scenarios
 - *investigating history & culture* - integration of ChatGPT into VR enables virtual visit to historical places and cultural landmarks
 - *development of soft skills* - practice and hone soft skills, *e.g.*, leadership, teamwork & communication through VR simulations augmented by ChatGPT
 - *extracurricular activities*
 - personalized learning, gamification of education, international cooperation, educator empowerment
- *VR & ChatGPT integration opens up new training and educational opportunities!*

AI Research

AI research race gets crazy

- practically impossible to follow all developments announced everyday
 - new announcement and publication of important work everyday!
- *industry leads research - academia lags behind*
 - trend observed even before 2015
- everyone excited to show off their work to the world
 - conference and github.com
 - biggest driving force behind unprecedented scale and speed of advancement of AI together with massive investment of capitalists



AI progress within a month - March, 2024

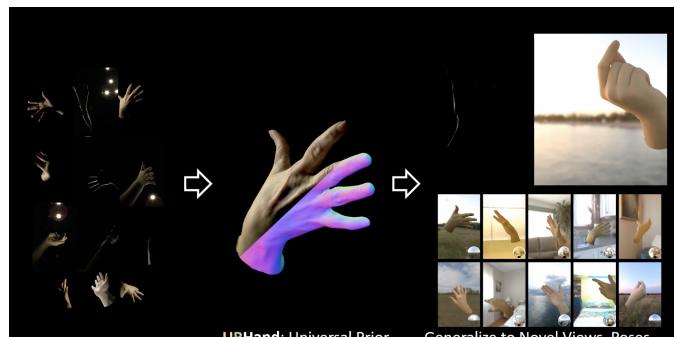
- UBTECH Humanoid Robot Walker S: Workstation Assistant in EV Production Line
- H1 Development of dance function
- Robot Foundation Models (Large Behavior Models) by Toyota Research Institute (TRI)
- Apple Vision Pro for Robotics
- Figure AI & OpenAI
- Human modeling
- LimX Dynamics' Biped Robot P1 Conquers the Wild Based on Reinforcement Learning
- HumanoidBench: Simulated Humanoid Benchmark for Whole-Body Locomotion and Manipulation - UC Berkeley & Yonsei Univ.
- Vision-Language-Action Generative World Model
- RFM-1 - Giving robots human-like reasoning capabilities

Papers of single company accepted by single conference

- CVPR 2024

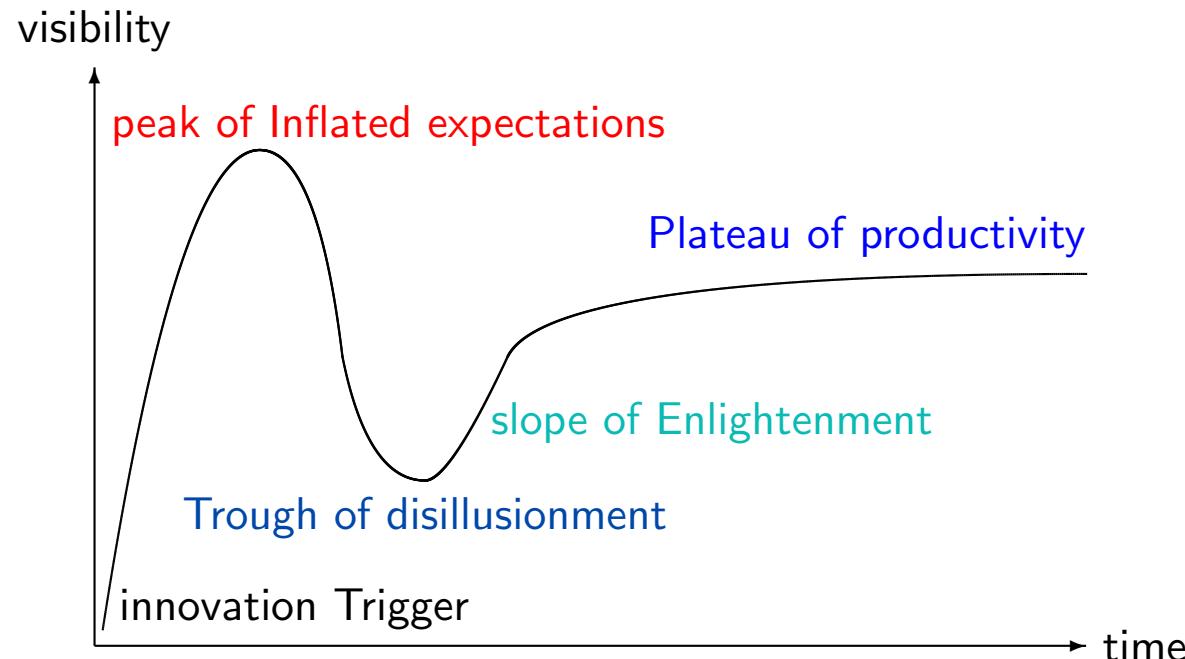


- PlatoNeRF: 3D Reconstruction in Plato's Cave via Single-View Two-Bounce Lidar - MIT, Codec Avatars Lab, & Meta [Klinghoffer et al., 2024]
 - 3D reconstruction from single-view
- Nymeria Dataset
 - large-scale multimodal egocentric dataset for full-body motion understanding
- Relightable Gaussian Codec Avatars - Codec Avatars Lab & Meta [Saito et al., 2024]
 - build high-fidelity relightable head avatars being animated to generate novel expressions
- Robust Human Motion Reconstruction via Diffusion (RoHM) - ETH Zürich & Reality Labs Research, Meta [Zhang et al., 2024]
 - robust 3D human motion reconstruction from monocular RGB videos



AI Market

AI hype cycle



- innovation trigger - technology breakthrough kicks things off
- peak of inflated expectations - early publicity induces many successes followed by even more
- trough of disillusionment - expectations wane as technology producers shake out or fail
- slope of enlightenment - benefit enterprise, technology better understood, more enterprises fund pilots

genAI products

- DALL-E (OpenAI)
 - trained on a diverse range of images
 - *generate unique and detailed images based on textual descriptions*
 - understanding context and relationships between words

- Midjourney
 - let people *create imaginative artistic images*
 - can interactively guide the generative process, providing high-level directions



genAI products

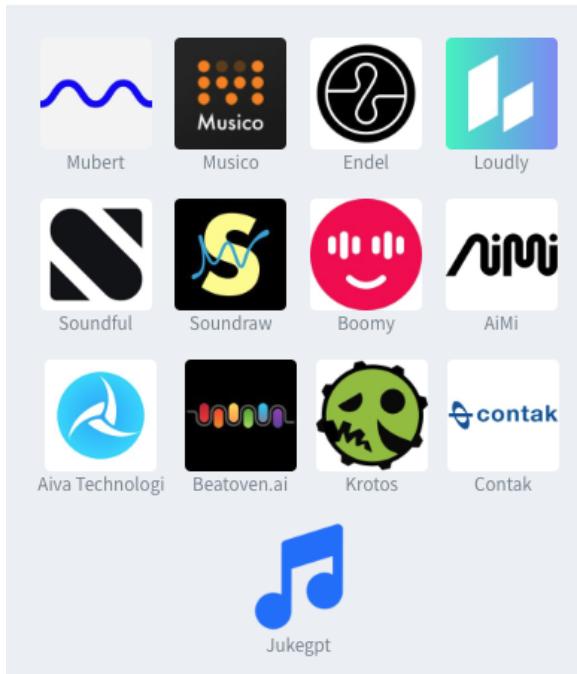


- Dream Studio
 - enables people to create music
 - *analyze patterns in music data and generates novel compositions based on input and style*
 - *allows musicians to explore new ideas and enhance their creative processes*
 - offer open-source free version
- Runway
 - provide range of generative AI tools for creative professionals
 - *realistic images, manipulate photos, create 3D models, automate filmmaking, . . .*
 - “artificial intelligence brings automation at every scale, introducing dramatic changes in how we create”

AI products

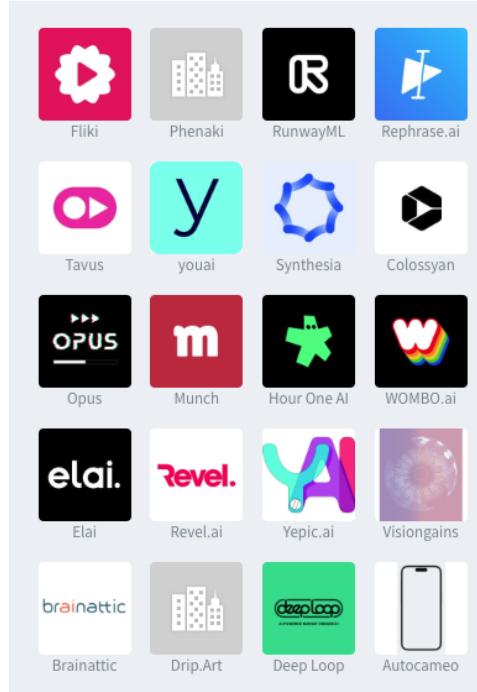
Audio: music generation

Combined funding \$ 61M



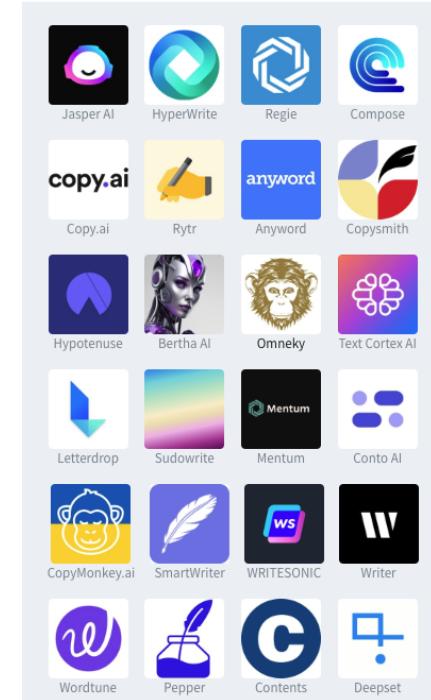
Video

Combined funding \$ 428M



Text: copy & writing

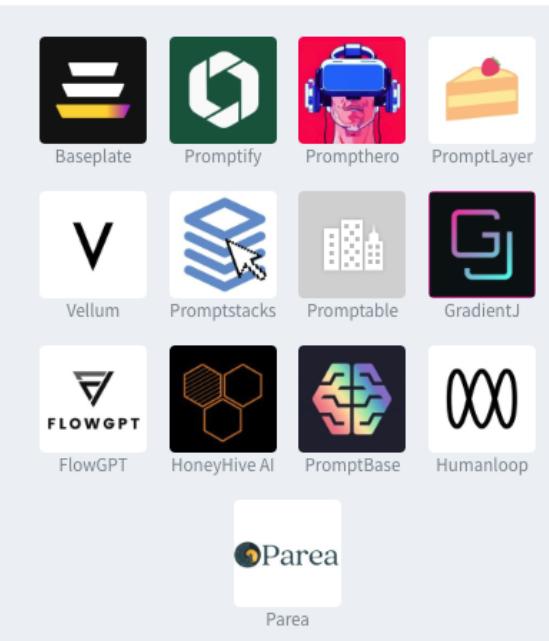
Combined funding \$ 863M



AI products

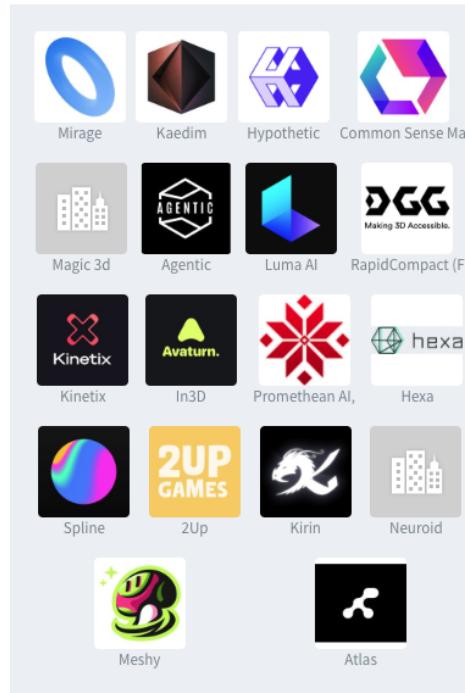
LLMs tools: Prompt Engineering and Management

Combined funding \$ 7.5M



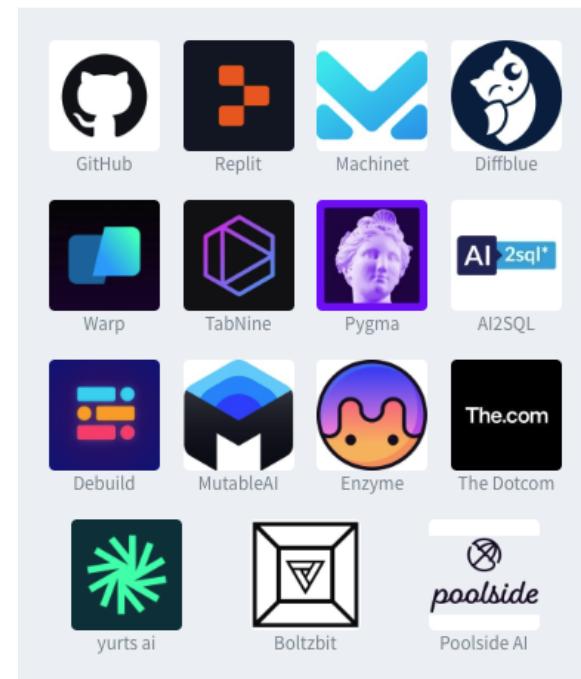
Gaming & design: 3d assets & worlds

Combined funding \$ 117M



Code: code generation

Combined funding \$ 828M



AI companies

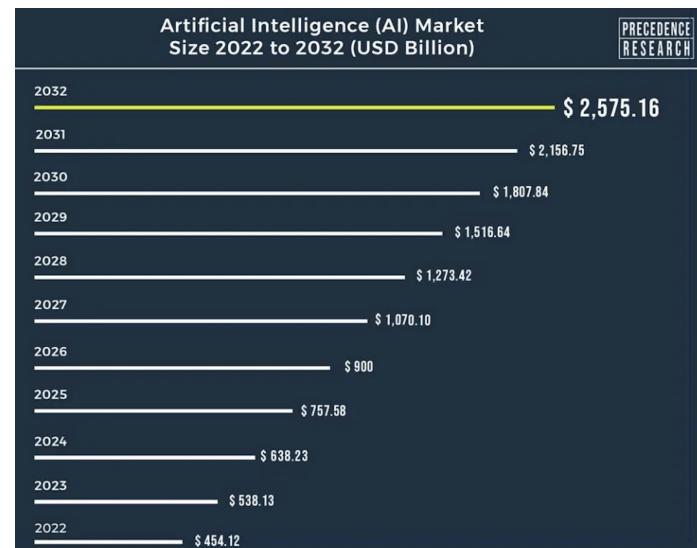
- big tech companies
 - OpenAI, Microsoft, Google, Meta - foundation models
- small(er) players
 - Figure AI, Mistral AI
- AI hardware companies - benefiting from LLM and genAI market dominance
 - Nvidia, AMD, Samsung, SK hynix, Micron, Intel, TSMC (AI processors & memory chips)
- *tiny fraction of Silicon Valley startups gets majority of total funding*
 - Anthropic - \$3.5B - large-scale AI systems - Claude
 - AssemblyAI - \$58M - speech AI
 - Hugging Face - \$400M - AI model/data platforms
 - Inflection AI - \$1.5B - conversational AI - Pi

Opportunities among big tech's domination

- OpenAI/Microsoft, Meta, Google's races for foundation models heated up!
- no small players can compete with rare exceptions, *e.g.*, Mistral AI
- hyperscalers stand strong - AWS, Azure, and Google Cloud
- *speaker's proposals for strategies*
 - accurately (or roughly) predict how far & up to where big players will reach
 - target niche markets
 - focus on (creative) downstream applications of LLMs and/or genAIs

AI market outlook in 2024

- global AI market expected to reach *USD 0.5T by 2024* (IDC @ Mar-2023) & expected to reach around *USD 2.5T by 2032* (Precedence Research @ Dec-2023) [P.R., 2023]
 - was valued at USD 454B in 2022, expanding at *double-digit CAGR of 19%* from 2023 to 2032
- *AI funding soars to USD 17.9B for Q3 in 2023 in Silicon Valley while rest of tech slumps* (PitchBook data, Bloomberg @ Oct-2023) [Bloomberg, 2023]
 - multibillion-dollar investment in AI startups almost commonplace in Silicon Valley
 - genAI dazzles users and investors with photo-realistic images & human-sounding text
- genAI software sales could surge *18,647% by 2032*



Productivity, inflation & jobs

- Federal Reserve probes AI's impact on productivity, inflation & jobs - Jul-2024 [AnalyticsInsight, 2024]
 - feds acknowledging significant AI investments
 - Jerome Powell emphasizes uncertainties on whether AI will eliminate, augment, or create jobs - stating it's too early to predict
 - Powell acknowledges limited influence of central banks like the Fed on AI's technological shifts
 - fed actively researching various AI forms beyond genAI to understand potential economic impacts
 - IMF predicts AI (could) impact up to 60% of jobs in advanced economies potentially lowering labor demand and wages in sectors like finance and insurance

AI & global economy

- five ways AI is transforming global economy [AnalyticsInsight, 2024]
 - reshape job markets, creating new roles while rendering some obsolete
 - enhance productivity across industries
 - contribute to global economy by optimizing processes and innovation
 - *may widen economic disparities if not managed inclusively*
 - *governments* has to develop policies to address AI's economic and social impacts



Global Semiconductor Markets

Hard-to-predict AI hardware markets

- US traditionally has strong design houses
 - Nvidia, Apple, . . . , Amazon, Google, . . .
- threatened by vulnerable supply chains experienced in COVID period → reshoring
- NOW *want to make chips themselves! - can and will reshape AI hardware industry*
- Intel declares seriousness about foundry business!
- Nvidia challenged!
 - many companies including AMD starting share AI chips markets
 - big techs start making their own hardware

Turmoils in global semiconductor market

- US CHIPS for America Act - semiconductor manufacturing reshoring
 - ask (or coerce) world-best semiconductor companies build factories in US with support of government and states
- Biden administration - US government - awards
 - \$1.5B @ Feb-2024 - Global Foundry
 - \$0.685B @ Apr-2024 - SK Hynix @ Lafayette, Indiana (Silicon Heartland) - next-generation memory chips for AI investing \$4B
 - \$6.4B @ Apr-2024 - Samsung @ Talor, Texas - chips for automotive, consumer technology, IoT, & aerospace investing \$40B
 - \$6.6B @ Apr-2024 - TSMC @ Phoenix, Arizona - Foundry
 - \$50M funding - small biz research and development
- TSMC's presence in Japan - backed by government

Case study - AMD - Nvidia's new competitor

- Instint MI300X - launched on 06-Dec-2023
 - 50% more HBM3 capacity than its predecessor, MI250X (128 GB)
 - *outperform Nvidia's H100 TensorRT-LLM* (when using optimized AI software stack)
 - 1.6X Higher Memory Bandwidth - 1.3X FP16 TFLOPS
 - up to 40% faster vs H100 (Llama-2 70B) in 8v8 server
- *great timing when Nvidia's order backlogs stuck*
- AMD stocks soars as of Jan-2024
- potential risks: ROCm vs CUDA, speed of customer adoption, production coverage



Serendipities around Als

Serendipity or inevitability

- What if Geoffrey Hinton had not been persistent researcher?
- What if symbolists won AI race over connectionists?
- What if attention mechanism did not perform well?
- What if Transformer architecture did not perform super well?
- What if Jensen Hwang had not been crazy about making hardware for professional gamers?
- Is it like Alexander Fleming's Penicillin?
- Or more like Inevitability?

Some Important Questions

Some important questions around AI

- why human-level AI in the first place?
- what lies in very core of DL architecture? what makes it work amazingly well?
- biases that can hurt judgement, decision making, social good?
- ethical and legal issues
- consciousness, knowledge, belief, reasoning
- future of AI

Human-level AI?

Why human-level in the first place?

- lots of times, when we measure AI performance, we say
 - how can we achieve human-level performance, *e.g.*, CV models?
- why human-level?
 - are all human traits desirable? are humans flawless?
 - aren't humans still evolving?
- advantage of AI over humans
 - *e.g.*, self-driving cars can use extra eyes, GPS, computer network
 - *e.g.*, recommendation system runs for hundreds of millions of people overnight
 - AI is available 24 / 7 while humans cannot
 - . . . critical advantages for medical assistance, emergency handling
 - AI does not make more mistakes because task is repetitive and tedious
 - AI does not request salary raise or go on strike

What makes DL so successful?

Factors contributing to astonishing success of DL

- analysis based on speaker's mathematical, numerical algorithmic & statistical perspectives considering hardware innovations

30% universal approximation theorem? - (partially) yes! but that's not all

- function space of neural network is *dense* (math theory), *i.e.*, for every $f : \mathbf{R}^n \rightarrow \mathbf{R}^m$, exists $\langle f_n \rangle$ such that $\lim_{n \rightarrow \infty} f_n = f$

25% architectures/algorithms tailored for each class of applications, *e.g.*, CNN, RNN, Transformer, NeRF, diffusion, GAN, VAE, . . .

20% data labeling - expensive, data availability - unlimited web text corpus

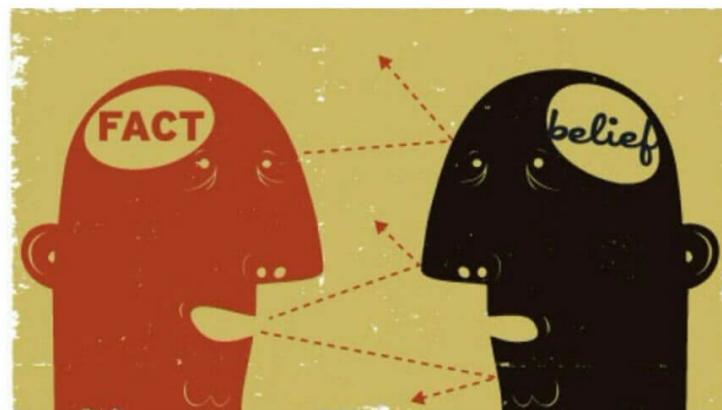
15% computation power/parallelism - AI accelerators, *e.g.*, GPU, TPU & NPU

10% rest - Python, open source software, cloud computing, MLOps, . . .

Biases - by Humans & Machines

Cognitive biases

- cognitive biases [Kahneman, 2011]
 - confirmation bias, availability bias
 - hindsight bias, confidence bias, optimistic bias
 - anchoring bias, halo effect, framing effect, outcome bias
 - belief bias, negativity bias, false consensus,



LLM biases

- plausible with LLM
 - availability bias - biased by imbalancedly available information
 - LLM trained by imbalanced # articles for specific topics
 - belief bias - derive conclusion not by reasoning, but by what it saw
 - LLM easily inferencing what it saw, *i.e.*, data it trained on
 - halo effect - overemphasize on what prestigious figures say
 - LLM trained by imbalanced # reports about prestigious figures
- similar facts true for other types of ML models,
 - *e.g.*, video caption, text summarization, sentiment analysis
- cognitive biases only human represent
 - confirmation bias, hindsight bias, confidence bias, optimistic bias, anchoring bias, negativity bias, framing effect

Ethical and Legal Issues

Ethics - possibilities & questions

- AI can be exploited by those who have bad intention to
 - manipulate / deceive people - using manipulated data corpus for training
 - *e.g.*, spread false facts
 - induce unfair social resource allocation
 - *e.g.*, medical insurance, taxation
 - exploit advantageous social and economic power
 - *e.g.*, unfair wealth allocation, mislead public opinion
- AI for Good - advocated by Andrew Ng
 - *e.g.*, public health, climate change, disaster management
- should scientists and engineers be morally & politically conscious?
 - *e.g.*, Manhattan project

Ethically controversial issues

- AI girlfriends
 - lots of AI girlfriend apps already developed
 - ethical considerations and provisions for user privacy with AI partners imperative - as with every technology involving personal data and emotional interaction
 - prospect of developing lifelike digital companions will grow better with evolution of AI
 - perhaps changing ways relationships and companionship perceived in digital age one day
 - why not many AI boyfriend apps? is this sexual discrimination issue (at all)?

Legal issues with ethical consideration - (hypothetical) scenarios

- scenario 1: full self-driving algorithm causes traffic accident killing people
 - who is responsible? - car maker, algorithm developer, driver, algorithm itself?
- scenario 2: self-driving cars kill less people than human drivers
 - e.g., human drivers kill 1.5 people for 100,000 miles & self-driving cars kill 0.2 people for 100,000 miles
 - how should law makers make regulations?
 - utilitarian & humanistic perspectives
- scenario 3: someone is not happy with their data being used for training
 - “The Times sues OpenAI and Microsoft over AI use of copyrighted work” (Dec. 2023)
 - “Newspaper publishers in California, Colorado, Illinois, Florida, Minnesota and New York said Microsoft and OpenAI used millions of articles without payment or permission to develop ChatGPT and other products” (Apr. 2024)

Consciousness

Consciousness

- what is consciousness, anyway?
 - recognizes itself as independent, autonomous, valuable entity?
 - recognizes itself as living being, unchangeable entity?
 - will to survive?
- no agreed definition on consciousness exists yet
 - ... and will be so forever
- can it be separated from fact that humans are biological living being?
 - (speaker) doesn't think so ...
- is SKYNET ever plausible (without someone's intention)?
 - can AI have *desire* to survive (or save earth)?



Utopia or dystopia



- not important questions (speaker thinks)
 - what we should worry about is not doomsday or destroying humankind
- but rather we should focus on
 - our limit in controlling or unintended consequences of AI
 - misuse by those possessing social, economic, political power
 - social good and welfare impaired by (exploiting of) AI
 - choice among utilitarianism / humanism / justice / equity
 - handle ethical and legal issues

Knowledge, Belief, and Reasoning of AI

Does LLM (or AI) have knowledge or belief? Can it reason?

What categories of questions should they be?

Philosophical? Cognitive scientific?

Three surprises of LLM

- LLM is very different sort of animal . . . except that it is *not* an animal!
- *unreasonable* effectiveness of data [Halevy et al., 2009]
 - *performance scales with size of training data*
 - *qualitative leaps* in capability as models scale
 - tasks demanding human intelligence *reduced to next token prediction*
- focus on third surprise
 - “*conditional probability model looks like human with intelligence*”
 - making vulnerable to anthropomorphism
- examine it by throwing questions
 - “*does LLM have knowledge and belief?*”
 - “*can it reason?*”

Knowledge, belief & reasoning around LLM

- *not* easy topic to discuss, or even impossible because
 - we do *not* have agreed definition of these terms especially in context of being asked questions like

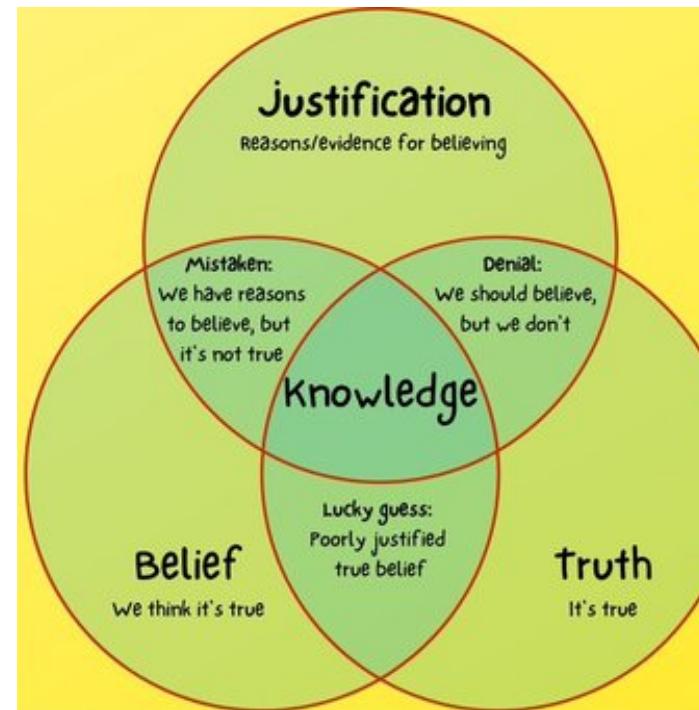
does ChatGPT have belief?
or
do humans have knowledge?
- let us discuss them in two different perspectives
 - laymen's perspective
 - cognitive scientific perspective

Laymen's perspective on knowledge, belief & reasoning

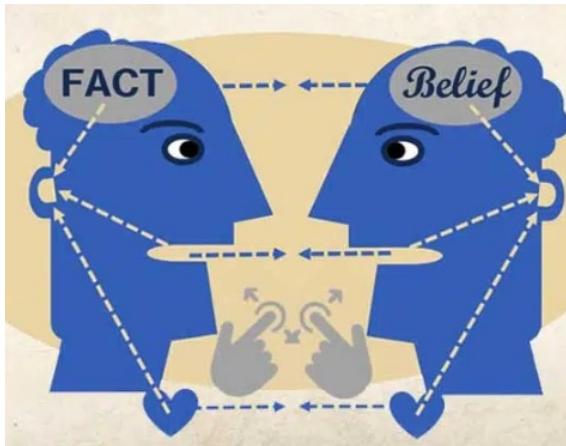
- does (good) LLM have knowledge?
 - Grandmother - looks like it cuz when instructed “*explaining big bang*”, it says
“The Big Bang theory is prevailing cosmological model that explains the origin and evolution of the universe. . . . 13.8 billion years ago . . .”
- does it have belief?
 - Grandmother: I don't think so, e.g., it does not believe in God.
- can it reason?
 - Grandmother: seems like it! e.g., when asked “*Sunghee is a superset of Alice and Beth is a superset of Sunghee. Is Beth a superset of Alice?*”, it says
“Yes, based on information provided, if Sunghee is a superset of Alice and Beth is a superset of Sunghee, then Beth is indeed a superset of Alice . . .”
- can it reason to prove theorem whose inferential structure is more complicated?
 - Grandmother: I'm not sure. - actually, I don't know what you're talking about!

Cognitive scientific perspective on knowledge

- does LLM have knowledge?
 - Speaker: I don't think so.
- why?
 - Speaker: we say we have "knowledge" when
"we do so against ground of various human capacities that we all take for granted when we engage in everyday conversation with each other."
 - LLM *cannot* do this.
 - Speaker: also when asked "*who is Tom Cruise's mother?*", it says "*Tom Cruise's mother is Mary Lee Pfeiffer.*" However, this is nothing but "*guessing*" by *conditional probability model* the *most likely* following words after "*Tom Cruise's mother is.*"
 - Speaker: so we *cannot say it really knows the fact!*



Cognitive scientific perspective on belief



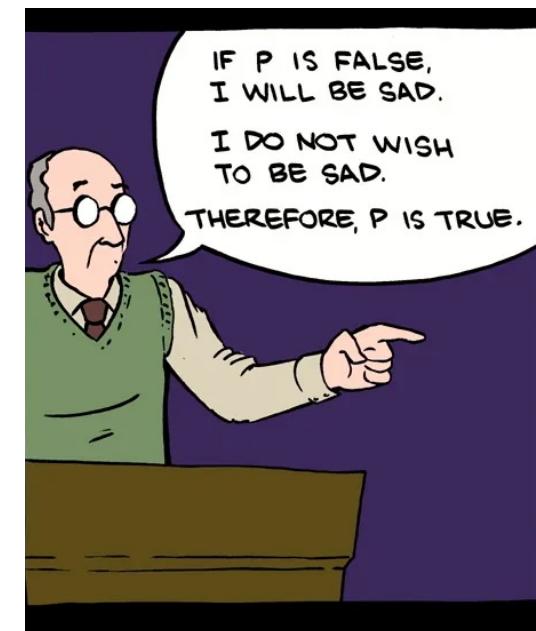
- for the discussion
 - we do not concern *any specific belief*
 - we concern prerequisites for ascribing any beliefs to AI system
- so does it have belief?
 - Speaker: nothing can count as belief about the world we share unless

it is against ground of the ability to update beliefs appropriately in light of evidence from that world, an essential aspect of the capacity to distinguish truth from falsehood.
 - Speaker: when a human being takes to Wikipedia and confirms some fact, what happens is not her language model update, but

reflection of her nature as language-using animal inhabiting shared world with a community of other language-users.
 - Speaker: LLM does not have this ground, an essential consideration when deciding whether it *really* had beliefs.
 - Speaker: so *no, LLM cannot have belief!*

Cognitive scientific perspective on reasoning

- note reasoning is *content neutral*
 - e.g., following logic is perfect regardless of truth of premises
if Socrates is a human and humen are immortal, then Socrates would have survived today.
- Speaker: when asked “*if humans are immortal, would Socrates have survived today?*”, LLM says
 - “*... it's logical to conclude that Socrates would likely still be alive today. ...*”
 - however, remember, once again, what we just asked it to do is *not* “deductive inference”, but
given the statistical distribution of words in public corpus, what words are likely to follow the sequence, “humans are immortal and Socreates is human therefore.”
- Speaker: so LLM cannot or rather does not reason
- however, LLM can *mimic even multi-step reasoning whose inferencing structure is complicated* using *in-context learning* or *few-short prompting!*



A simple example supporting reasoning incapability



- You

"Who is Tom Cruise's mother?"

- ChatGPT

"Tom Cruise's mother is Mary Lee Pfeiffer. She was born Mary Lee South. . . . Information about his family, including his parents, has been publicly available,"

- You

"Who is Mary Lee Pfeiffer's son?"

- ChatGPT

"As of my last knowledge update in January 2022, I don't have specific information about Mary Lee Pfeiffer or her family, including her son. . . ."

Future of AI

Aschenbrenner's essay

- Leopold Aschenbrenner, who left OpenAI showing concerns about safety, wrote *epic 165-page treatise* - Jun-2024
 - rapid progress
 - AI development (is) accelerating at unprecedented rate, predicting by 2027, AI models lead to intelligence explosion surpassing human intelligence
 - economic and security implications
 - trillions of dollars being invested into infrastructure supporting AI systems
 - critical need for securing technologies to prevent misuse, *e.g.*, by state actors like Chinese Communist Party (CCP)
 - technical and ethical challenges
 - significant challenges in controlling AI (smarter than humans), *i.e.*, “superalignment” problem, to prevent catastrophic outcomes
 - predictions and societal impact
 - few people truly understand scale of change by AI
 - potential for AI to reshape industries, enhance national security
 - pose new ethical and governance challenges

More about Aschenbrenner's essay

- AGI by 2027
 - seen AI advancing from preschool-level to high-schooler abilities in 4 years highlighting rapid progress from GPT-2 to GPT-4
- superintelligence following AGI - post AGI
 - rapid advancement from human-level to superhuman capabilities
- G-dollar investment on AI clusters
- national & global security dynamics
 - may lead to all-out war, *e.g.*, with China, if not managed properly
- superalignment challenges
 - keeping superintelligent AI aligned with human values and interests - “one of the most critical predictions”
- societal and economic transformations, project involvement by US government, technological mobilization

Moral

Moral

- AI, *e.g.*, LLM, shows incredible utility and commercial potentials, hence we should
 - make informed decisions about trustworthiness and safety
 - avoid ascribing capacities they lack
- today's AI is so powerful, so (seemingly) convincingly intelligent
 - obfuscate mechanism
 - actively encourage *anthropomorphism* with philosophically loaded words like “believe” and “think”
 - easily mislead people about character and capabilities of AI
- matters not only to scientists, engineers, developers, and entrepreneurs, but also
 - *general public, policy makers, media people*

Recent AI Research and New Development

Notable recent AI research and new development

- Kolmogorov–Arnold networks (KAN)
- JEPA (*e.g.*, I-JEPA & V-JEPA) & consistency-diversity-realism trade-off

KAN

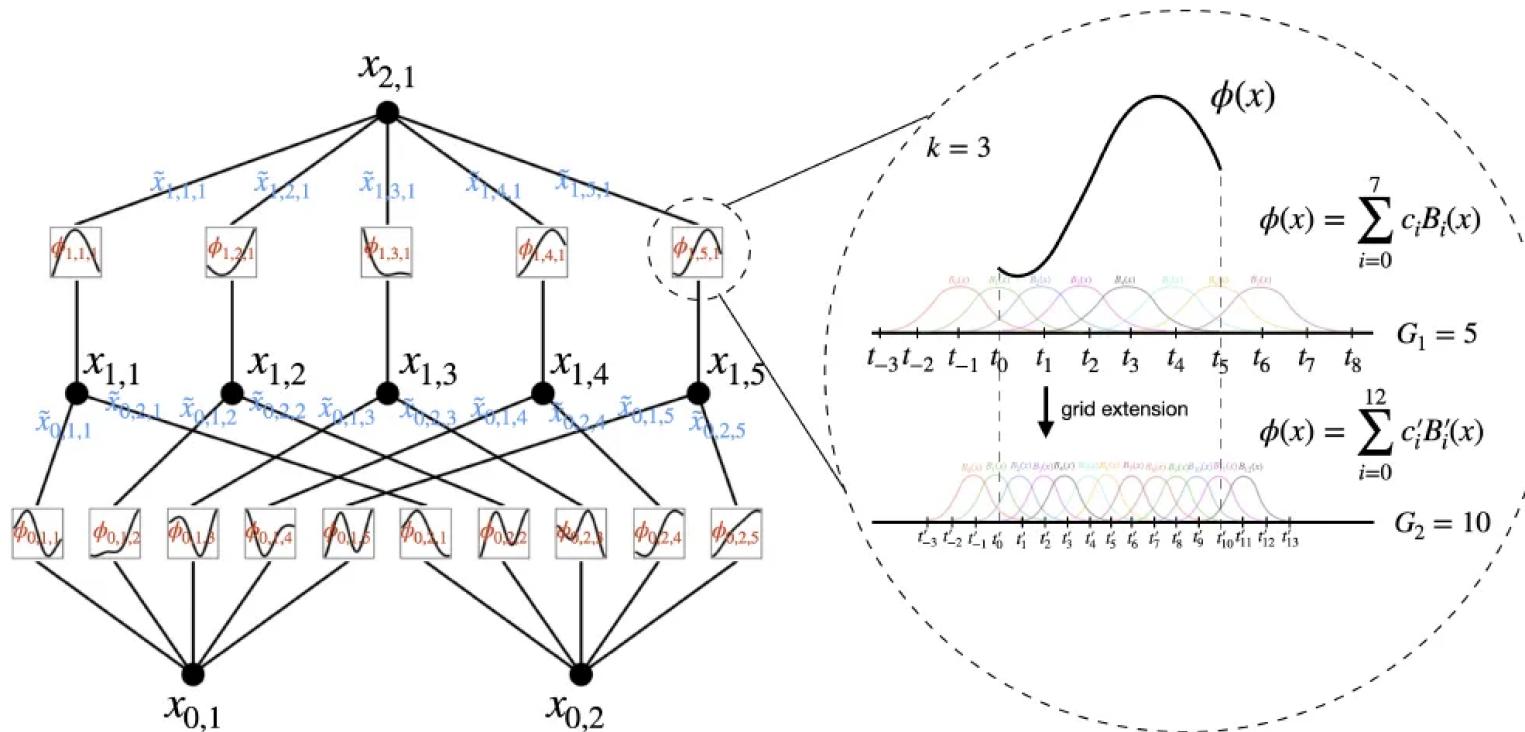
Kolmogorov–Arnold networks (KAN)

- KAN: Kolmogorov-Arnold Networks - MIT, CalTech, Northeastern Univ. & IAIFI
- techniques
 - inspired by Kolmogorov-Arnold representation theorem - every $f : \mathbf{R}^n \rightarrow \mathbf{R}$ can be written as finite composition of continuous functions of single variable, *i.e.*
 - $$f(x) = \sum_{q=0}^{2n} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$$
 - where $\phi_{q,p} : [0, 1] \rightarrow \mathbf{R}$ & $\Phi_q : \mathbf{R} \rightarrow \mathbf{R}$
 - replace (fixed) activation functions with learnable functions
 - use B-splines for learnable (uni-variate) functions - for flexibility & adaptability
- advantages
 - benefits structure of MLP on outside & splines on inside
 - reduce complexity and # parameters to achieve accurate modeling
 - *interpretable* by its nature
 - *better continual learning* - adapt to new data without forgetting thanks to local nature of spline functions

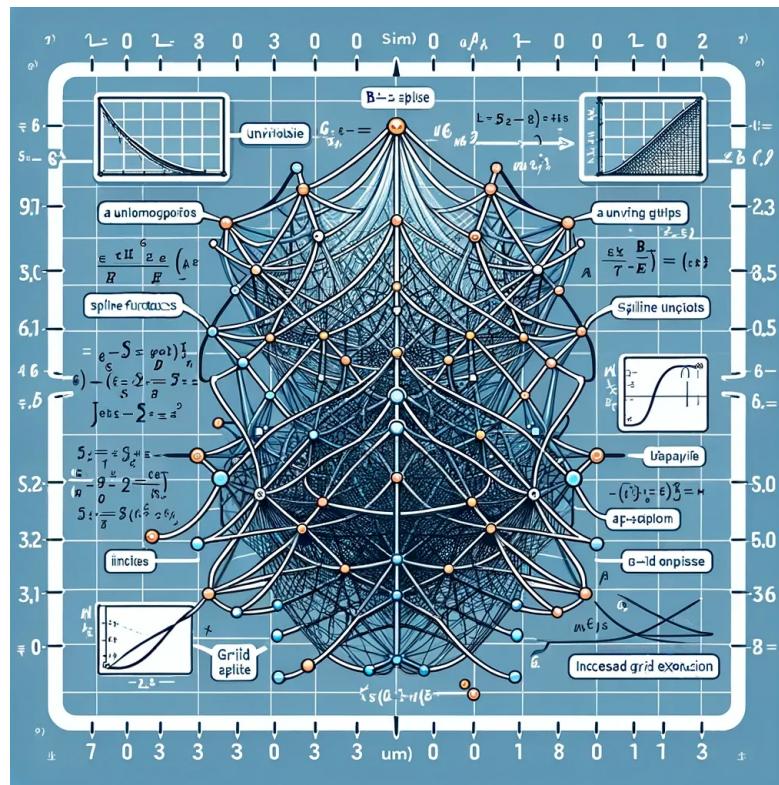
MLP vs KAN

Model	Multi-Layer Perceptron (MLP)	Kolmogorov-Arnold Network (KAN)
Theorem	Universal Approximation Theorem	Kolmogorov-Arnold Representation Theorem
Formula (Shallow)	$f(\mathbf{x}) \approx \sum_{i=1}^{N(e)} a_i \sigma(\mathbf{w}_i \cdot \mathbf{x} + b_i)$	$f(\mathbf{x}) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^n \phi_{q,p}(x_p) \right)$
Model (Shallow)	<p>(a)</p> <p>fixed activation functions on nodes</p> <p>learnable weights on edges</p>	<p>(b)</p> <p>learnable activation functions on edges</p> <p>sum operation on nodes</p>
Formula (Deep)	$\text{MLP}(\mathbf{x}) = (\mathbf{W}_3 \circ \sigma_2 \circ \mathbf{W}_2 \circ \sigma_1 \circ \mathbf{W}_1)(\mathbf{x})$	$\text{KAN}(\mathbf{x}) = (\Phi_3 \circ \Phi_2 \circ \Phi_1)(\mathbf{x})$
Model (Deep)	<p>(c)</p> <p>MLP(\mathbf{x})</p> <p>\mathbf{W}_3</p> <p>σ_2</p> <p>\mathbf{W}_2</p> <p>σ_1</p> <p>\mathbf{W}_1</p> <p>\mathbf{x}</p> <p>nonlinear; fixed</p> <p>linear; learnable</p>	<p>(d)</p> <p>KAN(\mathbf{x})</p> <p>Φ_3</p> <p>Φ_2</p> <p>Φ_1</p> <p>\mathbf{x}</p> <p>nonlinear; learnable</p>

KAN architecture with spline parametrization unit layer



Future work on KAN



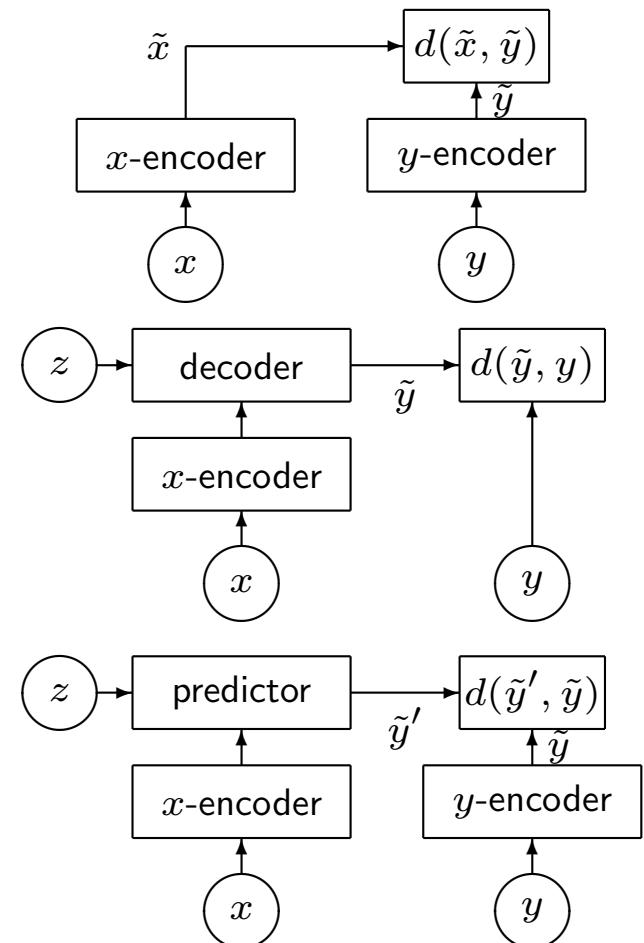
- natural question is
 - what if use both MLP and KAN?
 - what if use other types of splines?
 - how to control forgetfulness of continual learning?
 - why functions of one variable? possible to use functions of two variables?

(figure created by DALLE-3)

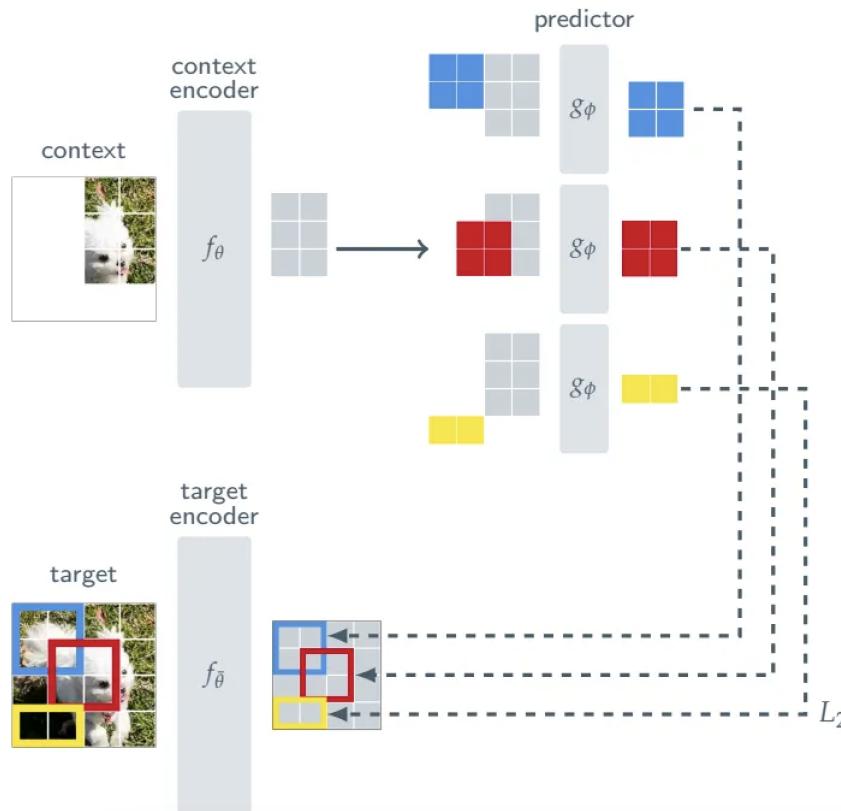
JEPA

Joint-Embedding Predictive Architecture (JEPA)

- Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture (JEPA) - Yann LeCun et al. - Jan-2023
 - joint-embedding architecture (JEA)
 - output similar embeddings for compatible inputs x, y and dissimilar embeddings for incompatible inputs
 - generative architecture
 - directly reconstruct signal y from compatible signal x using decoder network conditioned on additional variables z to facilitate reconstruction
 - joint-embedding predictive architecture (JEPA)
 - similar to generative architecture, but comparison is done in embedding space
 - e.g., I-JEPA learns y (masked portion) from x (unmasked portion) conditioned on z (position of mask)



Learning semantic representation better



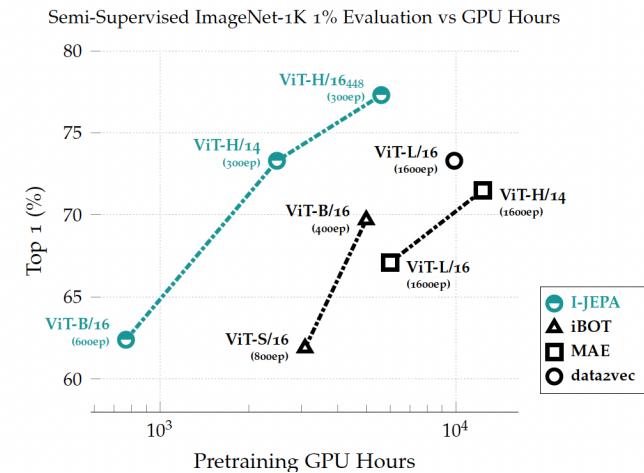
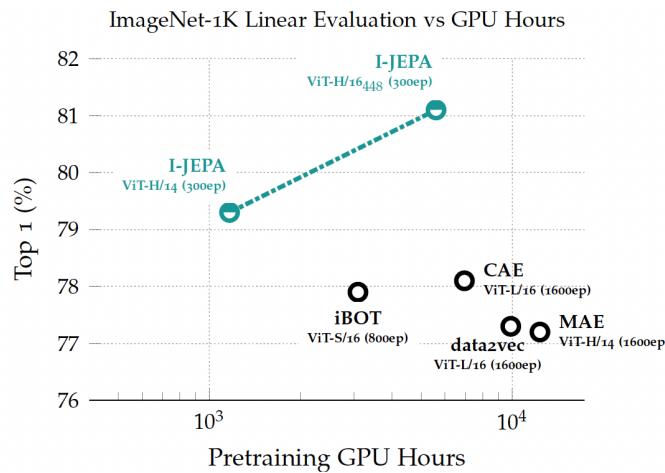
- I-JEPA

- predicts missing information in *abstract representation space*
- e.g., given single context block (unmasked part of the image), predict representations of various target blocks (masked regions of same image) where target representations computed by learned target-encoder
- generates semantic representations (not pixel-wise information) potentially eliminating unnecessary pixel-level details & allowing model to concentrate on learning more semantic features

I-JEPA outperforms other algorithms

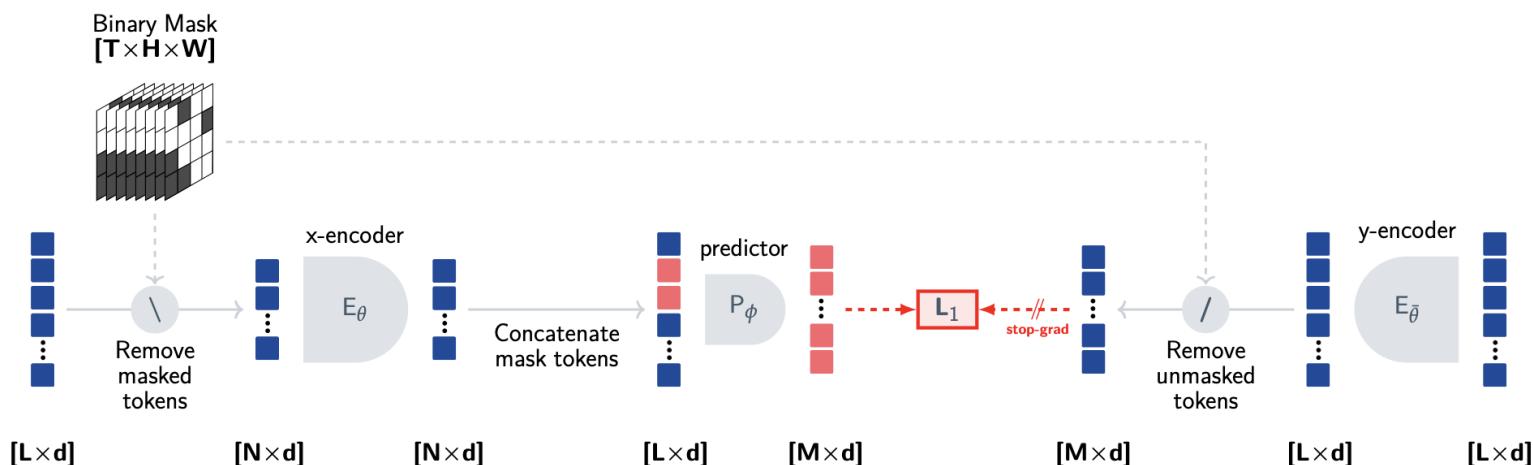
Method	Arch.	CIFAR100	Places205	iNat18
<i>Methods without view data augmentations</i>				
data2vec [8]	ViT-L/16	81.6	54.6	28.1
MAE [36]	ViT-H/14	77.3	55.0	32.9
I-JEPA	ViT-H/14	87.5	58.4	47.6
<i>Methods using extra view data augmentations</i>				
DINO [18]	ViT-B/8	84.9	57.9	55.9
iBOT [79]	ViT-L/16	88.3	60.4	57.3

Method	Arch.	Clevr/Count	Clevr/Dist
<i>Methods without view data augmentations</i>			
data2vec [8]	ViT-L/16	85.3	71.3
MAE [36]	ViT-H/14	90.5	72.4
I-JEPA	ViT-H/14	86.7	72.4
<i>Methods using extra data augmentations</i>			
DINO [18]	ViT-B/8	86.6	53.4
iBOT [79]	ViT-L/16	85.7	62.8



V-JEPA

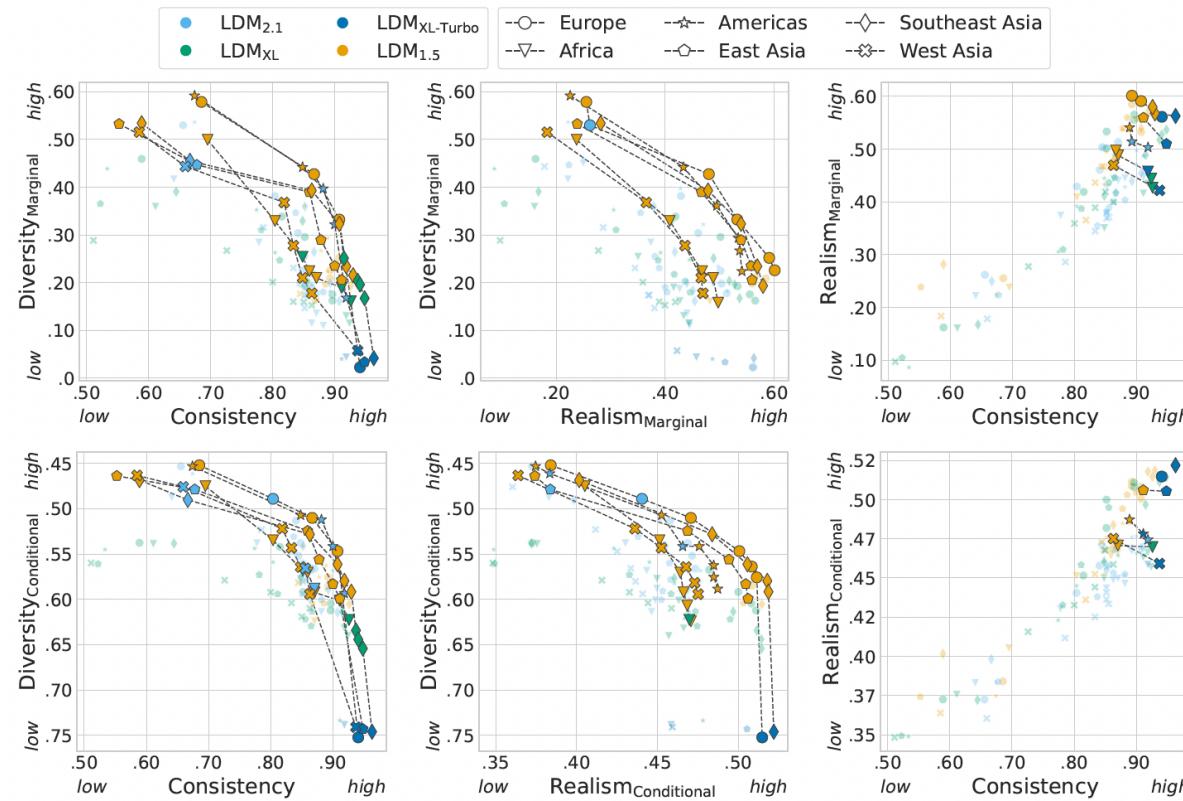
- Revisiting Feature Prediction for Learning Visual Representations from Video - Yann LeCun et al. - Feb-2024
 - essentially same ideas of JEPA - loss function is calculated in embedding space - for better semantic representation learning (rather than pixel-wise learning)



More realistic generative model becomes, less diverse it becomes

- Consistency-diversity-realism Pareto fronts of conditional image generative models - FAIR at Meta - Montreal, Paris & New York City labs, McGill University, Mila, Quebec AI institute, Canada CIFAR AI - Jun-2024
 - realism comes at the cost of coverage, *i.e.*, *the most realistic systems are mode-collapsed!*
 - intuition (or hunch)
 - world models should *not* be generative - should make predictions in representation space - in representation space, unpredictable or irrelevant information is absent
- main argument in favor of JEPA

Consistency-diversity-realism trade-off



References & Informants

References

- [AnalyticsInsight, 2024] AnalyticsInsight (2024). Analytics insight.
- [Assran et al., 2023] Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture.
- [Astolfi et al., 2024] Astolfi, P., Careil, M., Hall, M., Mañas, O., Muckley, M., Verbeek, J., Soriano, A. R., and Drozdzal, M. (2024). Consistency-diversity-realism pareto fronts of conditional image generative models.
- [Bardes et al., 2024] Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., and Ballas, N. (2024). Revisiting feature prediction for learning visual representations from video.
- [Bloomberg, 2023] Bloomberg (2023). Bloomberg.
- [Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

- [Goodfellow et al., 2014] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks.
- [Halevy et al., 2009] Halevy, A., Norvig, P., and Fernando, N. (2009). The unreasonable effectiveness of data. *Intelligent Systems, IEEE*, 24:8 – 12.
- [Kahneman, 2011] Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York.
- [Kingma and Welling, 2019] Kingma, D. P. and Welling, M. (2019). An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392.
- [Klinghoffer et al., 2024] Klinghoffer, T., Xiang, X., Somasundaram, S., Fan, Y., Richardt, C., Raskar, R., and Ranjan, R. (2024). Platonerf: 3d reconstruction in plato's cave via single-view two-bounce lidar.
- [Liu et al., 2024] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., and Tegmark, M. (2024). KAN: Kolmogorov-arnold networks.
- [Morency et al., 2022] Morency, L.-P., Liang, P. P., and Zadeh, A. (2022). Tutorial on multimodal machine learning. In Ballesteros, M., Tsvetkov, Y., and Alm, C. O., editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 33–38, Seattle, United States. Association for Computational Linguistics.

[P.R., 2023] P.R. (2023). Precedence research.

[Saito et al., 2024] Saito, S., Schwartz, G., Simon, T., Li, J., and Nam, G. (2024). Relightable gaussian codec avatars.

[Shanahan, 2023] Shanahan, M. (2023). Talking about large language models.

[Vaswani et al., 2023] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

[Yin et al., 2024] Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. (2024). A survey on multimodal large language models.

[Zhang et al., 2024] Zhang, S., Bhatnagar, B. L., Xu, Y., Winkler, A., Kadlecak, P., Tang, S., and Bogo, F. (2024). Rohm: Robust human motion reconstruction via diffusion.

Selected references & informants

- Daniel Kahneman, Thinking, Fast and Slow, 2011
- S. Yin, et. al., A Survey on Multimodal LLMs, 2023
- M. Shanahan, Talking About Large Language Models, 2022
 - M. Shanahan - Professor of *Cognitive Robotics* at Imperial College London
- D.P. Kingma, M. Welling. Introduction to Variational Autoencoders, 2019
- A. Vaswani, et al., Attention is all you need, NeurIPS, 2017
- I.J. Goodfellow, . . . , Y. Bengio, Generative adversarial networks (GAN), 2014
- A.Y. Halevy, P. Norvig, and F. Pereira. Unreasonable Effectiveness of Data, 2009
- Stanford Venture Investment Groups
- CEOs & CTOs @ startup companies in Silicon Valley
- VCs on Sand Hill Road - Palo Alto, Menlo Park, Woodside in California

END of SLIDES