DGIST Seminar
LLM & genAI - Technology, Business Applications, and Some Important Questions

Sunghee Yun
sunghee.yun@gmail.com

VP - AI Technology & Product Strategy
Erudio Bio, Inc.

Speaker

- VP - AI Technology & Product Strategy @ Erudio Bio, Inc., CA, USA
- Advisory Professor, Electrical Engineering and Computer Science @ DGIST, Korea
- Adjunct Professor, Electronic Engineering Department @ Sogang University, Seoul
- Technology Consultant @ Gerson Lehrman Gruop (GLG), NYC, USA
- CTO - Senior Fellow @ Gauss Labs Inc., Palo Alto, CA, USA ~ 2023

- Senior Applied Scientist @ Amazon.com, Inc., Vancouver, BC, Canada ~ 2020
- Principal Engineer @ Software R&D Center of Samsung DS Division, Korea ~ 2017
- Principal Engineer @ Strategic Marketing Team of Memory Business Unit, Korea ~ 2016
- Principal Engineer @ Memory Design Technology Team of DRAM Development Lab., Korea ~ 2015
- Senior Engineer @ CAE Team of Samsung Semiconductor, Korea ~ 2012
- M.S. & Ph.D. - Electrical Engineering (EE) @ Stanford University ~ 2004
- B.S. - Electrical Engineering (EE) @ Seoul National University ~ 1998

Takeaways and questions

- purpose of this talk is to answer questions such as
  - what are LLM & genAI and how do they work?
  - what is the secret sauce behind LLM?
  - how AI-equipped products will change our lives?
  - what will AI business and market of 2024 look like?

- another purpose is to *make the audience curious* about topics like
  - what are the things that we should be careful about when using AI systems?
  - how can we prevent potential harms while utilizing remarkable capabilities of AI?
  - how can / should we prepare for coming changes by AI?
  - (philosophical) questions like . . . is AI intelligent? knowledgable? biased?

So today we will discuss

- large language model (LLM)
  - definition & examples
  - multi-modality
  - attention turns out to be way more *crucial*
    . . . than even original authors envisioned!
  - technicality around Transformer
- generative AI (genAI)
  - definition, some models, and applications
- business applications
  - applications, companies, and products
- AI market trend, 2024 outlook, and startup strategies
- some important topics & questions around & future of AI
  - why human-level performance?
  - AI ethics, law, biases, consciousness

Large language model (LLM)

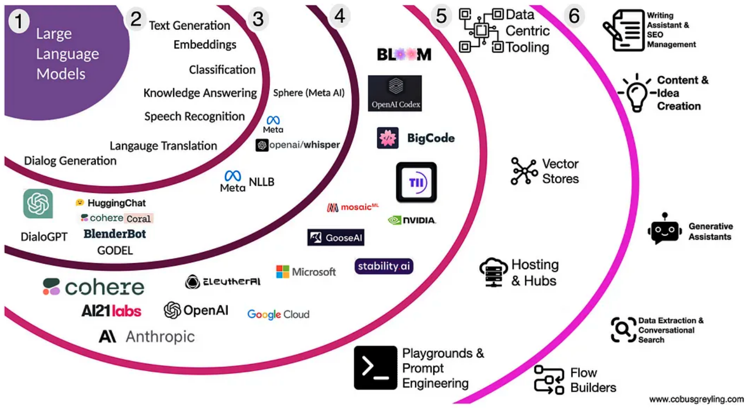
- LLM
  - is a type of AI aimed for natural language processing (NLP) trained on massive corpus of texts (and programming code)
  - allows learn statistical relationships between words & phrases, $i.e.$, conditional probability
  - surprise everyone - unreasonable effectiveness of data (Halevry et al, 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

History of NLP

- bag of words - first introduced in 1954
- word embedding in 1980
- recurrent neural networks (RNN) based models
- long shot-term memory (LSTM) - based on RNN - in 1997
- 380M-sized seq2seq model using LSTMs proposed in 2014
- 130M-sized seq2seq model using gated recurrent units (GRUs)
- Transformer in 2017 - Attention is All You Need (by A. Vaswani)
  - 100M-sized encoder-decoder multi-head attention model
  - remove recurrent architecture, handling arbitrarily long dependencies
  - parallelizable
  - simple linear-transformation-based attention model

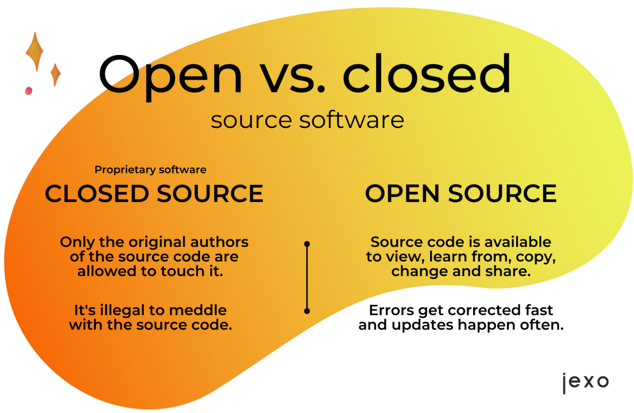
LLM examples

- foundation models
  - GPT-x (deriving Chat-GPT) - OpenAI, Llamax - Meta, PaLMx (Bard) - Google
- # parameters
  - generative pre-trained transfomer (GPT) - GPT-1: 117 M, GPT-2: 1.5 B, GPT-3: 175 B, GPT-4: 100 T
  - large language model Meta AI (Llama) - Llama1: 65 B, Llama2: 70 B
  - scaling language modeling with pathways (PaLM) - 540 B
- trained with huge corpus burning a lot of cash on GPUs!
- applied to many NLP & multi-modal genAI applications
- enabled by efficiently parallelizable attention mechanism - Transformer!



LLM open source vs closed model

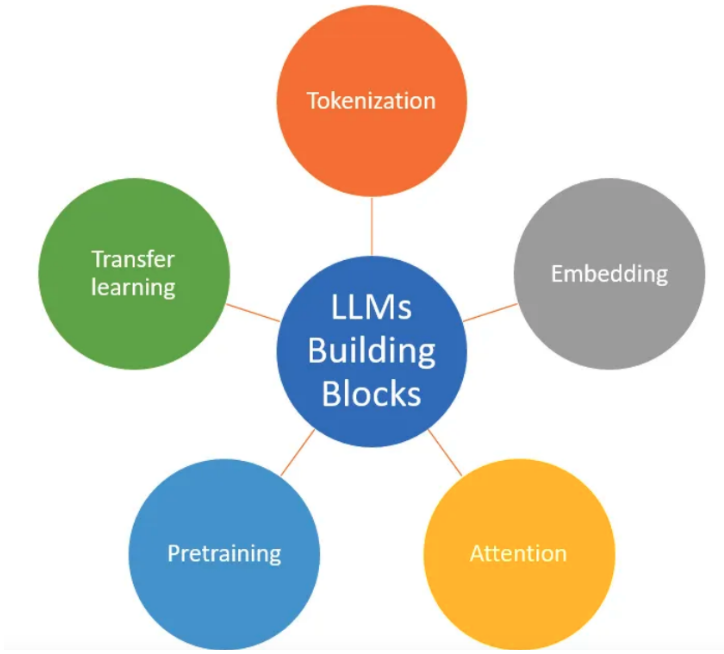
- open source
  - people can view, learn from, copy, change or share it
  - transparency, control, collaboration, training, security, stability, frequent release, inclusion, community
  - freedom to
    * run program
    * study / change source code
    * redistribute copies
    * distribute modified versions
  - examples: GPT-2, *Llama*, Hugging Face



- proprietary software / closed source
  - available to people, team, company that created it
  - examples: GPT-3.5, GPT-4, Bard

LLM building blocks

- data
  - trained on massive datasets of text & code
  - quality & size critical on performance
- architecture
  - can make huge difference
  - example: Mitral 7B - on par with Llama2 (70B)
- training
  - self-supervised learning
  - supervised learning, *e.g.*, reinforcement learning via human feedback (RLHF) by ChatGPT
- inference
  - LLM generates output, *e.g.*, content creation, text summarization
  - in-context learning, prompt engineering



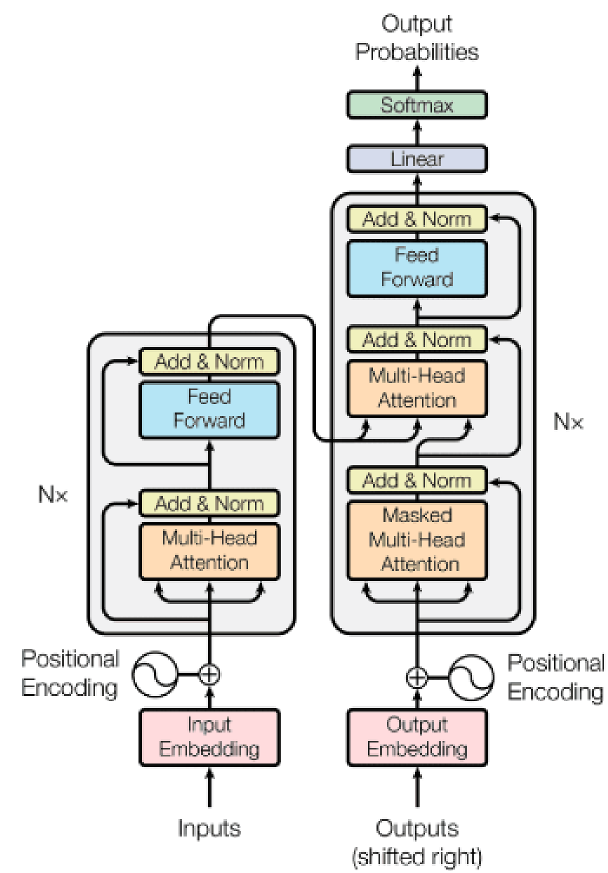
LLM architectural secret (or rather public) sauce

Transformer - parallelizable (simpler) attention mechanism

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

Transformer architecture

- encoding-decoding architecture
  - input embedding space $\rightarrow$ multi-head & mult-layer representation space $\rightarrow$ 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
- 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 $\rightarrow$ parallelizable $\rightarrow$ fast w/ additional cost in computation



Single-head scaled dot-product attention

Here values / keys / queries denote value / key / query *vectors* and d_k & d_v are lengths of keys / queries & vectors respectively

Also we use standard linear algebra notions for matrices and vectors - not transposed version that (almost) all ML scientists (wrongly) use

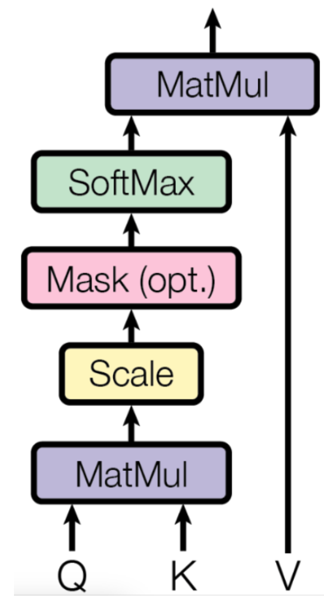
- output: weighted-average of values where weights are attentions / dependencies among tokens
- assume n queries and m key-value pairs

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

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

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

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



Single-head - close look at equations

- focus on ith query, $q_i \in \mathbf{R}^{d_k}$, $Q = \begin{bmatrix} - & q_i & - \end{bmatrix} \in \mathbf{R}^{d_k \times n}$
- assume m keys and m values, $k_1, \ldots, k_m \in \mathbf{R}^{d_k}$ & $v_1, \ldots, v_m \in \mathbf{R}^{d_v}$

$$K = \begin{bmatrix} k_1 & \cdots & k_m \end{bmatrix} \in \mathbf{R}^{d_k \times m}, V = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix} \in \mathbf{R}^{d_v \times m}$$

- 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 ith output token and jth 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 ith query, q_i in $\mathrm{Attention}(Q, K, V)$

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

Multi-head attention

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

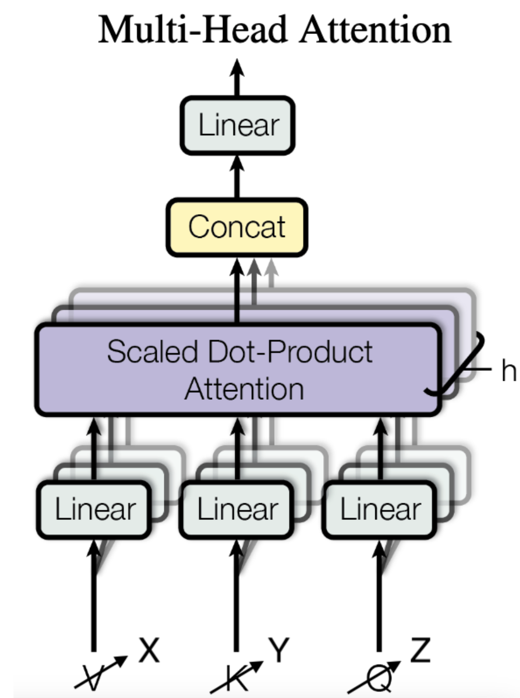
$$X \in \mathbf{R}^{d_e \times m}, \ Y \in \mathbf{R}^{d_e \times m}, \ Z \in \mathbf{R}^{d_e \times n}$$

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

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

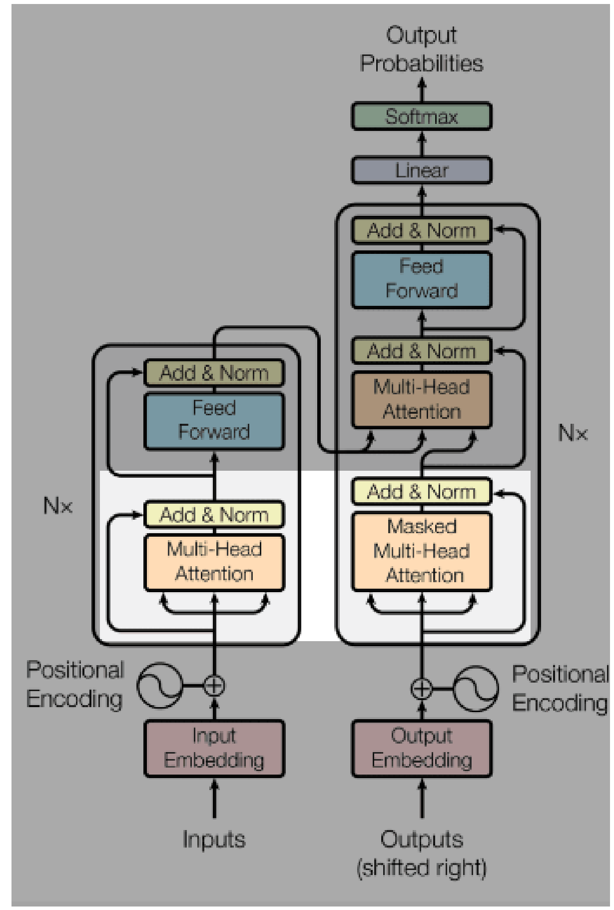
$$W^O \begin{bmatrix} A_1 \\ \vdots \\ A_h \end{bmatrix} \in \mathbf{R}^{d_e \times n},$$

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



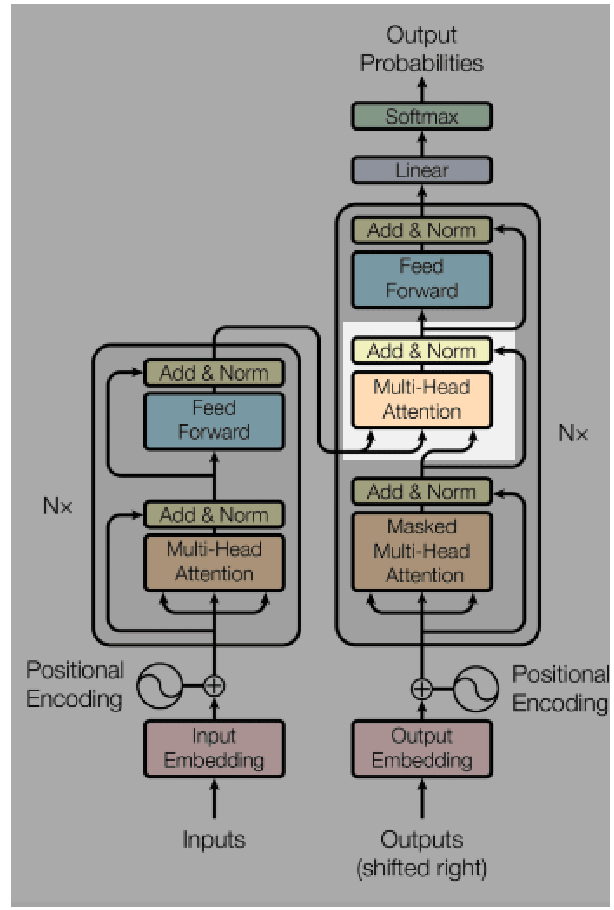
Self attention

- $m = n$
- 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

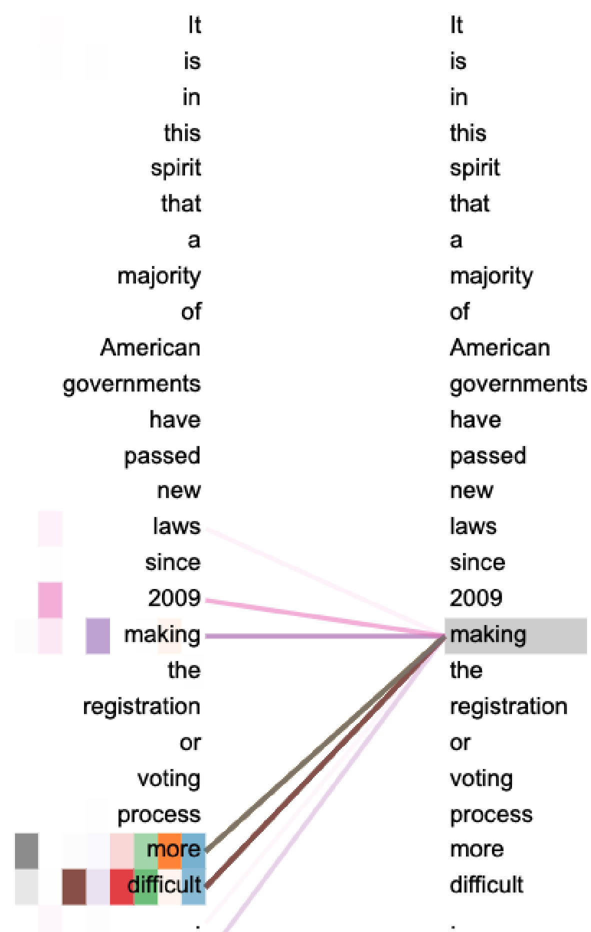


Visualization of self attentions - 1

example sentence:

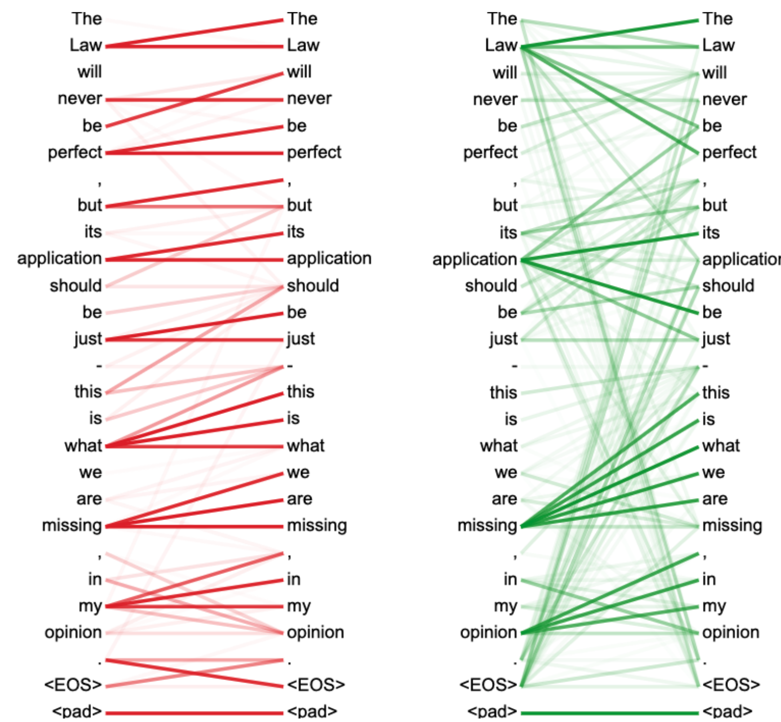
"It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult."

- 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"

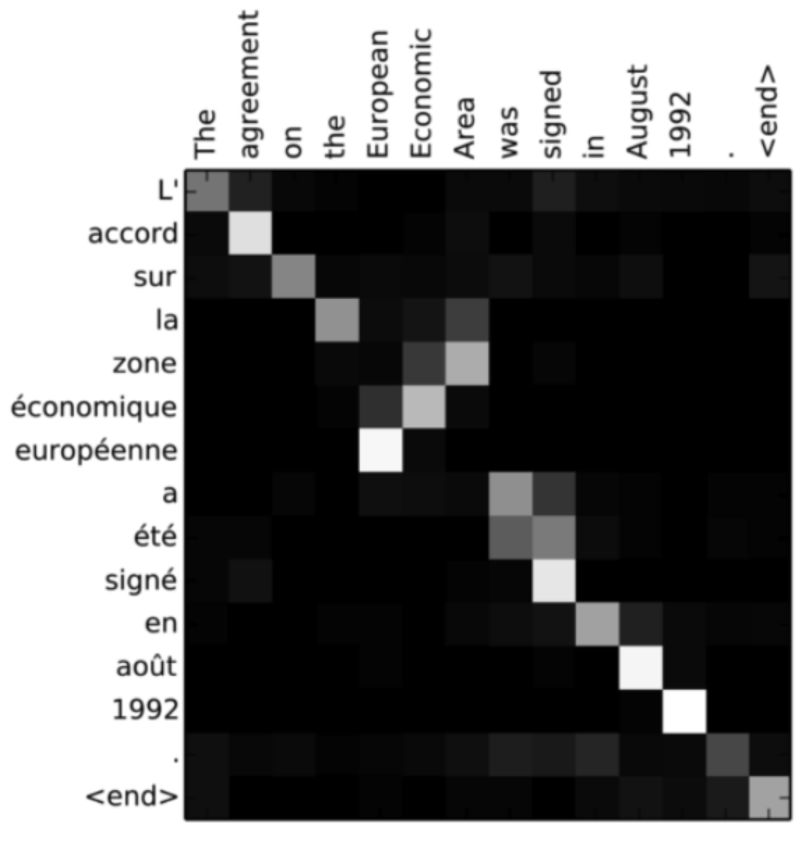


Visualization of self attentions - 2

- self attentions of encoder for two heads (of a layer)
  - different heads represent different structures
    → advantages of multiple heads
  - multiple heads work together to colletively 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 $\rightarrow$ 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."

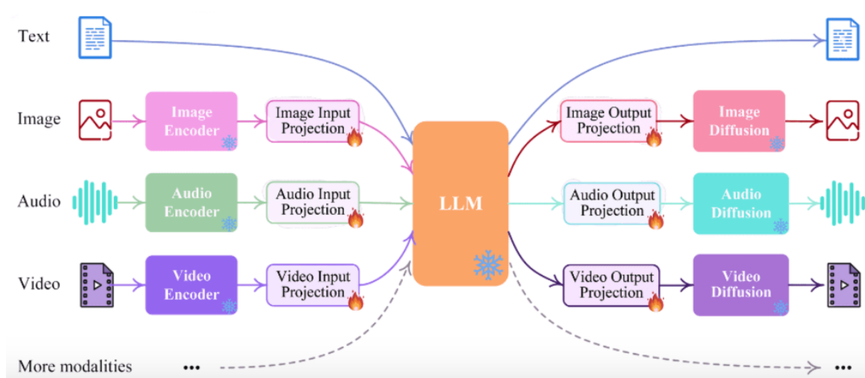
- encoder-decoder attention reveals relevance between
  - European $\leftrightarrow$ européenne
  - Economic $\leftrightarrow$ européconomique
  - Area $\leftrightarrow$ zone

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: $\mathcal{O}(n)$
- Transformer-based models
  - GPT-3, GPT-4, ChatGPT, BERT, Llama1, Llama2, PaLM

Multi-modality

- ability to understand information from multiple modalities, *e.g.*, text, images, audio, and video
- training methods
  - language representation + image / video / text / audio representation
  - learn multi-modal 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)



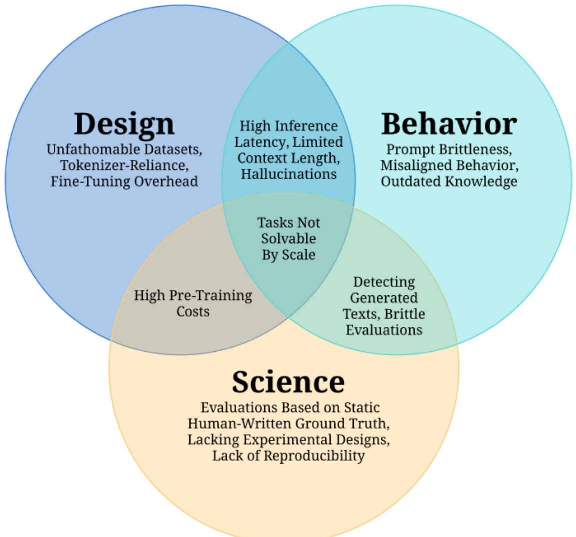
Research trends and directions (proposed)

- (very) many researchers change gears towards LLM
  - computer vision (CV), speach, music, video . . .
  - even reinforcement learning

- why LLM?
  - LLM is not necessarily about language!
  - LLM can connect non-NLP world using specific language structures
    $e.g.$, humans have handed down knowledge using natural languages for thousands of years
  - natural language evolved in ways optimized for knowledge transfer
  - interal representation structure of natural language optimized in such a way

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 iceburg" found & releaved

generative AI (genAI)

- definition of generative model

$$\boxed{\mathcal{Z}} \xrightarrow{g_\theta(z)} \boxed{\mathcal{X}}$$

  - generate samples in original space, $\mathcal{X}$, from samples in embedding (or 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 plausiable point in $\mathcal{X}$
  - inference: random samples z to generated target samples $x = g_\theta(z)$

    $e.g.$, image, text, voice, music, video

genAI early model - VAE

- variational auto-encoder (VAE)

$$\boxed{\mathcal{X}} \xrightarrow{q_\phi(z|x)} \boxed{\mathcal{Z}} \xrightarrow{p_\theta(x|z)} \boxed{\mathcal{X}}$$

  – log-likelihood: for any $q_\phi(z|x)$

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

  – (approximately) maximize likelihood by maximizing evidence lower bound (ELBO)

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

  – generative model: $p_\theta(x|z)$

genAI early model - GAN

- generative adversarial networks (GAN)

$$
\begin{tikzcd}
 & p(x) \arrow[drr, "x_{\mathrm{data}}"] & & \\
q(z) \arrow[r, "z"] & \boxed{g(\theta_G; z)} \arrow[r, "x_{\mathrm{model}}"] & \boxed{f(\theta_D; x)} \arrow[r] & \text{true / false}
\end{tikzcd}
$$

  - value function

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

  - 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

- maximze conditional probability

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

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

  - previous sequence: $x_{t-1}, x_{t-2}, \ldots$
  - 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

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

genAI applications

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

Industry genAI applications - 1

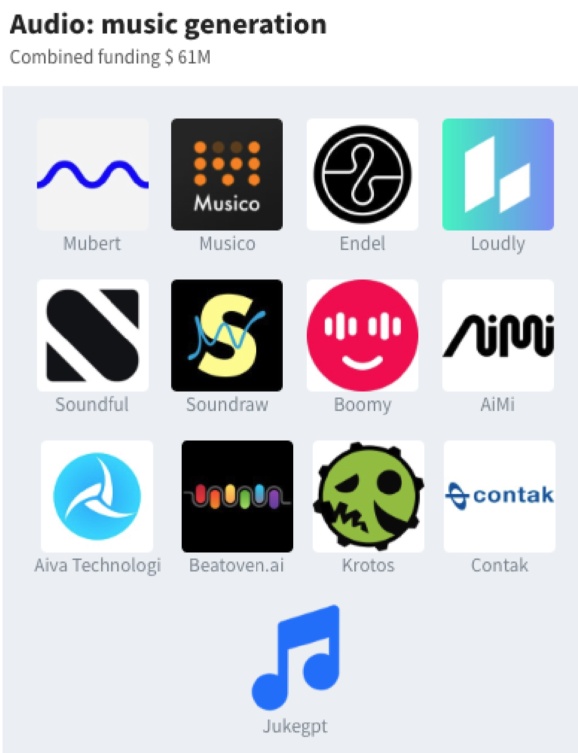
- DALL-E (OpenAI)
  - trained on a diverse range of images
  - generate unique and detailed images based on textual descriptions
  - understanding context and relationships between words

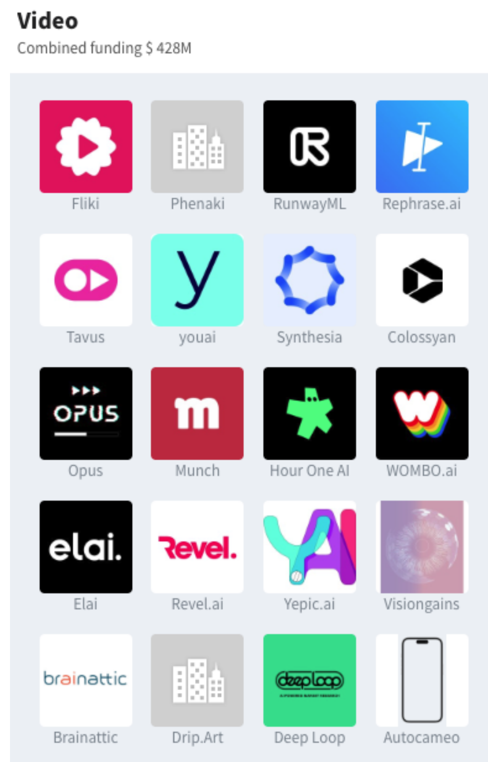
- Midjourney
  - let people create imaginative artistic images
  - can interactively guide the generative process, providing high-level directions

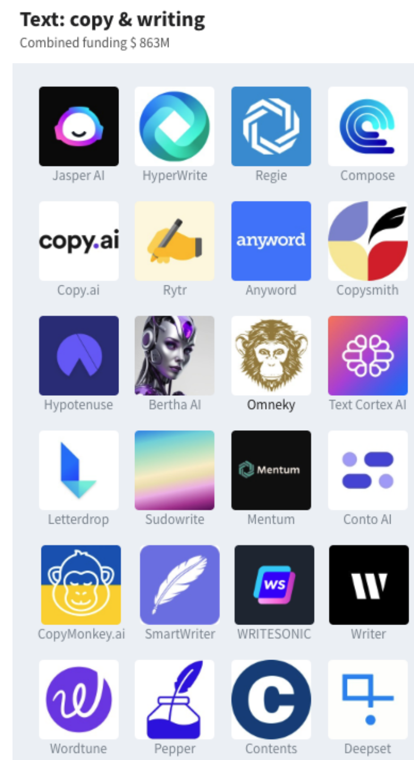
Industry genAI applications - 2

- 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"

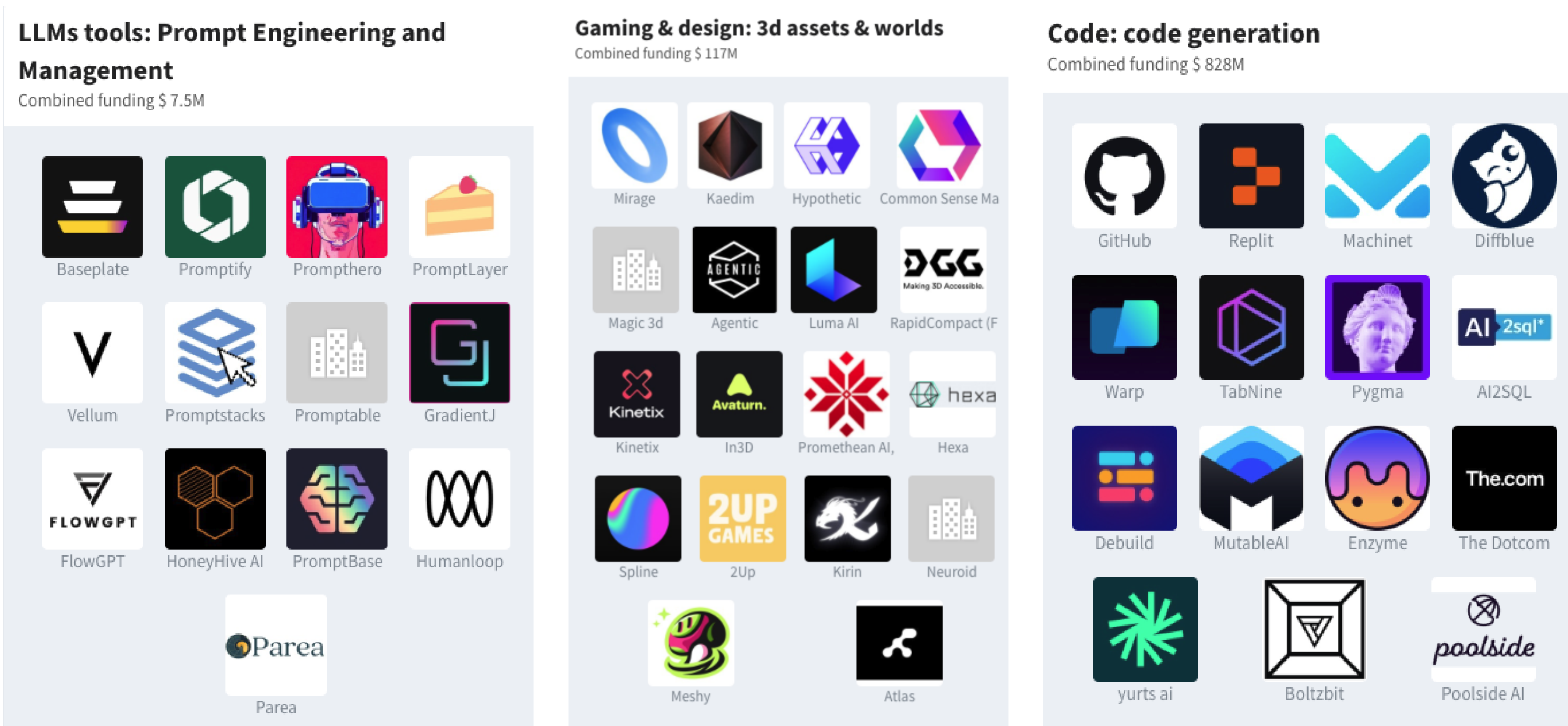
AI products - 1







AI products - 2

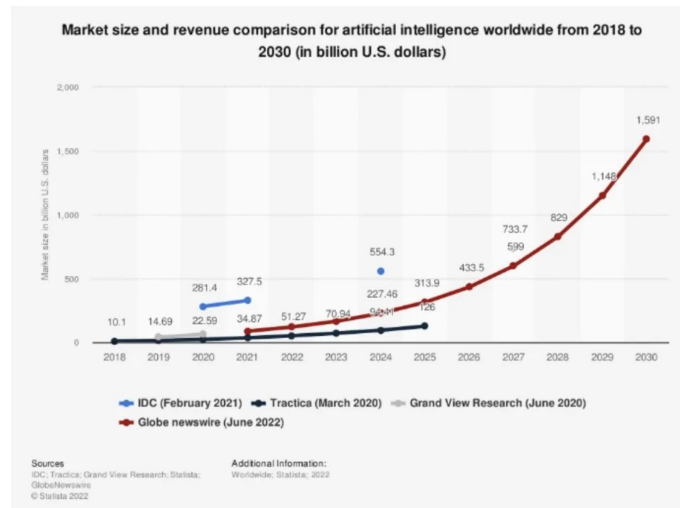


AI companies & products

- big players
  - Google, Meta, Microsoft, OpenAI, Mistral AI, Amazon (?) - foundation models
  - Nvidia, AMD, Samsung, SK hynix, Micron, Intel, TSMC - GPUs & memory chips
- (tiny fraction of) Silicon Valley startups - total funding
  - Anthropic - $3.5B - AI safe and research service
  - AssemblyAI - $58M - transcribe and understand speech
  - Hugging Face - $400M
  - Inflection AI - $1.5B
- startups by Korean founders
  - 12 labs - multi-modal video understanding & search
  - 24dot - self-driving car
  - Sapion, Rebellion - AI accelerator chips

AI market trends in 2024

- global AI market expected to reach \$0.5T by 2024 (by IDC, 15-Mar-2023, yahoo! finance)
- AI funding soars to \$17.9B for Q3 in 2023 in Silicon Valley while rest of tech slumps (by PitchBook data, 17-Oct-2023, Bloomberg)
  - multibillion-dollar investment in AI starupts almost commonplace in Silicon Valley
  - genAI dazzles users and investors with photo-realistic images & human-sounding text
  - web HTML moment for genAI
- but
  - other tech fell, *e.g.*, info tech hardware, healthcare, consumer goods
  - even AI less than post-pandemic peak in 2021
  - big deals for standout companies, *e.g.*, Anthropic & OpenAI dominates

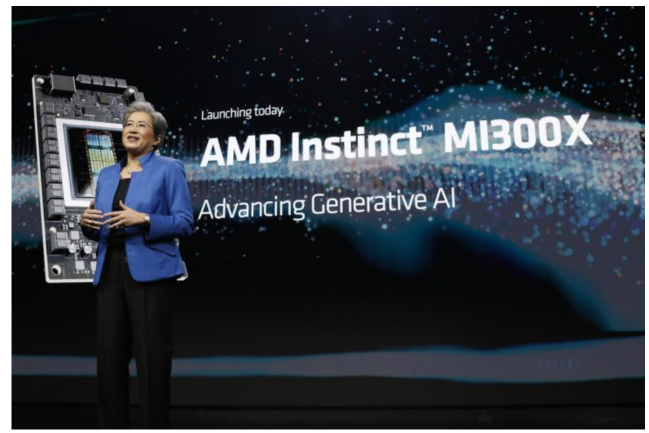


AI startup strategies for Silicon Valley in 2024

- prediction - by expert VCs
  - 2024 will be neither another 2021 nor guaranteed to improve
  - life-science looking good compared to other techs
- proposals
  - startups should steadily show values for following 3 – 5 years
  - prepared for risk hedge, pivoting, adoption of new biz models according to market and user attraction
  - not shoot for big around of investment, but incremental atraction
  - large startups utmostly effort to reduce expenses while shooting for many small investments

AMD - Nvidia's new competitor

- Instint MI300X - launched on 06-Dec-2023
  - designed solely on the CDNA 3 architecture, mix of 5nm and 6nm IPs, all combining to deliver 153B transistors
  - 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
- adopted by customer, LaminiAI backed by AMD
- great timing when Nvidia's order backlogs stuck
- AMD stocks soars as of Jan-2024
- Lisa Su (AMD's CEO) categorizes them as next big thing in tech industry
- potential risks: ROCm vs CUDA, unpredictable customer adoption



Some important questions around AI

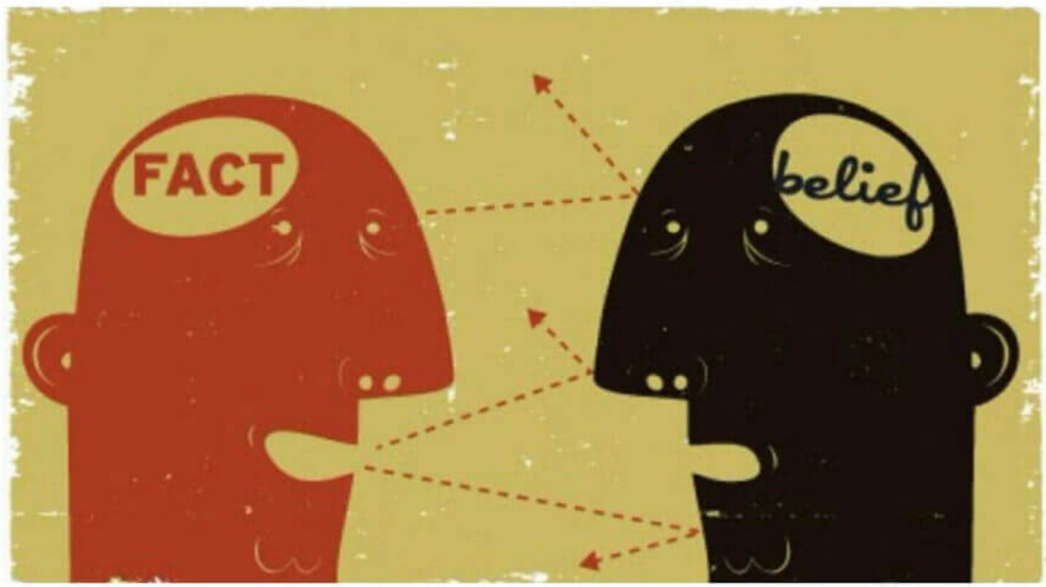
- why human-level AI in the first place?
- biases that can hurt judgement, decision making, social good?
- ethical and legal issues
- consciousness
  - can we even define it?
- contemplation on knowledge, belief, and reasoning around LLM
  - (and for that matter) around general AIs

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 advanages for medical assitance, emergency handling
  - AI does not make more mistakes because task is repetative and tedius
  - AI does not request salary raise or go on strike

Cognitive biases

- there exist biases such as
  - confirmation bias
  - availability bias
  - hindsight bias
  - confidence bias
  - optimistic bias
  - anchoring bias
  - belief bias
  - negativity bias
  - halo effect
  - framing effect
  - false consensus
  - outcome bias



LLM biases

- plausible with LLM
  - availability bias - baised 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 eaisly 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
  - false consensus - overemphasize how much others share their beliefs & values
    - LLM trained by comments by opinionated commenters
- 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

Ethics - possibilities & questions

- AI can be exploited by those who have bad intention to
  - manupilate / deceive people - using manupilated 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, $e.g.$
  - $e.g.$, public health, climate change, disaster management
- should scientists and engineers be morally & politically conscious?
  - $e.g.$, Manhattan project

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)

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

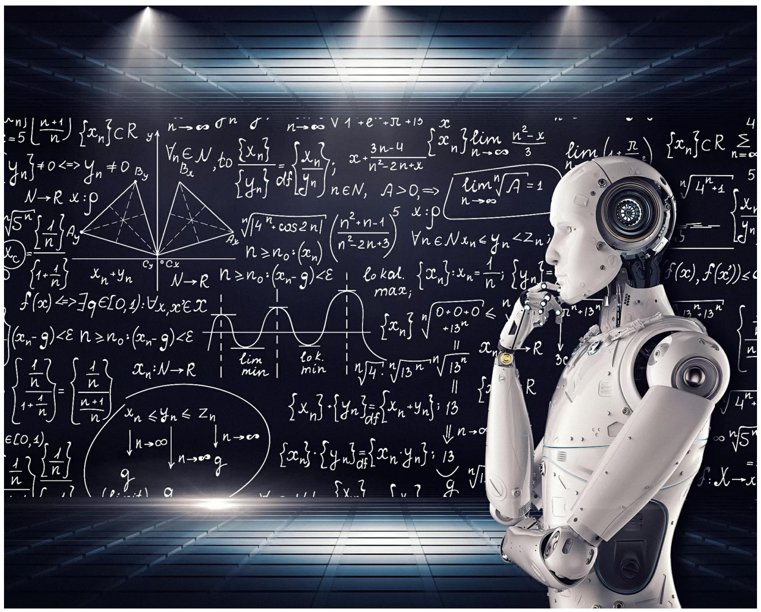
- can it be seperated from fact that humans are biological living being?
  - (speaker) doesn't think so . . .

- 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 doomday or destoying 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 welfair imparied by (exploting of) AI
  - choice among utilitarianism / humanism / justice / equity
  - handle ethical and legal issues

Other interesting questions



- *knowledge, belief, and reasoning of LLM/AI*
- is AI/LLM intelligent?
  - scientific perspective
  - brain scientific perspective
  - cognitive-scientific perspective
- impacts on labor and job market
  - reality / optimism / pessimism / resolution / prediction
- how should we prepare for our own futures

Does LLM have knowledge or belief? Can it reason?

Are they philosophical or cognitive scientific questions?

Or should they be some other types of questions?

Three surprises of LLM

- LLM is very different sort of animal . . . except that it is *not* an animal!
- *unreasonable* effectiveness of data (Halevry 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, and reasoning around LLM

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

does the GPT-4 have belief?

or

does a human have knowledge?

- we discuss them in two different perspectives
  - laymen's perspective
  - cognitive scientific perspective

Laymen's perspective on knowledge / belief / reasoning

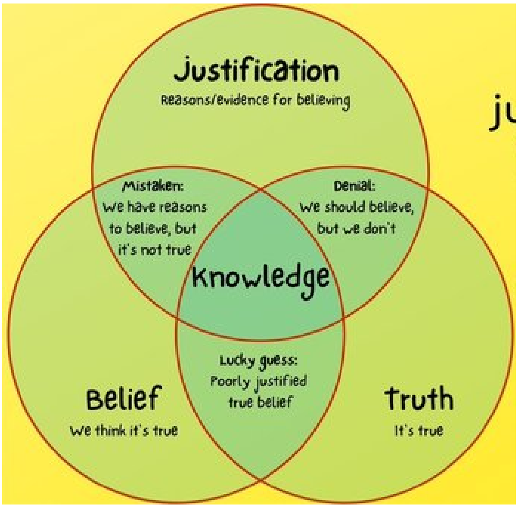
- does (a good) LLM have knowledge?
  - Grandmother: it looks like it, $e.g.$, when instructed "*explaing big bang*", ChatGPT says

    The Big Bang theory is the 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: it 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?"*, ChatGPT says

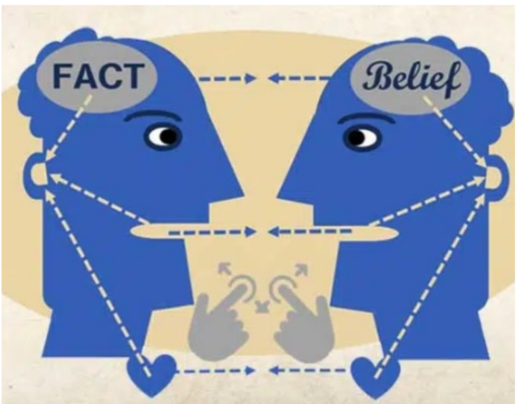
    Yes, based on the information provided, if Sunghee is a superset of Alice and Beth is a superset of Sunghee, then Beth is indeed a superset of Alice . . .
- can it reason to prove a theorem whose inferential structure is more complicated?
  - Grandmother: I'm not sure.

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?", ChatGPT 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!*



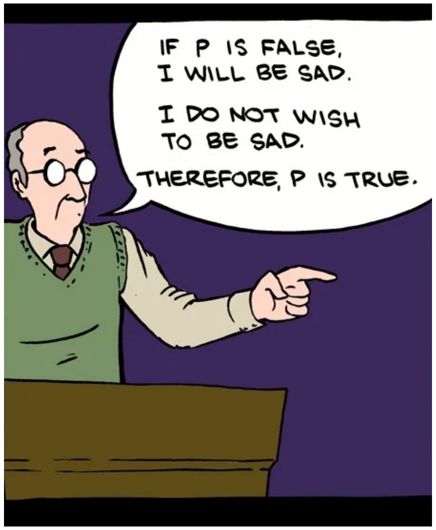
Cognitive scientific perspective on belief



- for the discussion
  - we do not concern *any specific belief.*
  - we concern the prerequisites for ascribing any beliefs to AI system.
- so does it have belief?
  - Speaker: nothing can count as a 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*
  - $e.g.$, the following logic is perfect regardless of truth of premises
    if Socrates is a human and humen are immortal, then Socrates would live today.
- Speaker: when asked "if humans are immortal, would Socrates live today?", ChatGPT 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.* . . .

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
  - take best usage of remarkable capabilities of AI

- 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

- this matters not only to scientists, engineers, developers, and entrepreneurs, but also
  - *general public, policy makers, media people*

References

- A.Y. Halevry, P. Norvig, and F. Pereira. The unreasonable effectiveness of data, 2009
- M. Shanahan, Talking About Large Language Models, 2022
- I.J. Goodfellow, . . . , Y. Bengio, Generative adversarial networks (GAN), 2014
- A. Vaswani, et al. Attention is all you need, NeurIPS, 2017
- D.P. Kingma, M. Welling. An Introduction to Variational Autoencoders, 2019
  - M. Shanahan - Professor of *Cognitive* Robotics at Imperial College London

Thank You!

Sunghee Yun

sunghee.yun@gmail.com