[San Jose State University Special AI Lecture Series II - LLM & GenAI Deep Dive] Inside the Generative Revolution - Transformers and Technology-Market Nexus of LLMs

Sunghee Yun

Co-Founder & CTO @ Erudio Bio, Inc.
Co-Founder & CEO @ Erudio Bio Korea, Inc.
Leader of Silicon Valley Privacy-Preserving AI Forum (K-PAI)
CGO / Global Managing Partner @ LULUMEDIC
Global Leadership Initiative Fellow @ Salzburg Global Seminar
Visiting Professor & Advisory Professor @ Sogang Univ. & DGIST

About Speaker

- *Co-Founder & CTO @ Erudio Bio, Inc., San Jose & Novato, CA, USA* 2023 ∼
- *Co-Founder & CEO @ Erudio Bio Korea, Inc., Korea* 2025 ∼
- *Leader of Silicon Valley Privacy-Preserving AI Forum (K-PAI), CA, USA* 2024 ∼
- *CGO / Global Managing Partner @ LULUMEDIC, Seoul, Korea* 2025 ∼
- *KFAS-Salzburg Global Leadership Fellow @ Salzburg Global Seminar, Austria* 2024 ∼
- *Adjunct Professor, EE Department @ Sogang University, Seoul, Korea* 2020 ∼
- *Advisory Professor, EECS Department @ DGIST, Korea* 2020 ∼
- *AI-Korean Medicine Integration Initiative Task Force Member @ The Association of Korean Medicine, Seoul, Korea* 2025 ∼
- *Director of AI Semiconductor @ K-BioX, CA, USA* 2025 ∼
- Global Advisory Board Member @ Innovative Future Brain-Inspired Intelligence System Semiconductor of Sogang University, Korea 2020 ∼
- Technology Consultant @ Gerson Lehrman Gruop (GLG), NY, USA 2022 ∼
- Chief Business Development Officer @ WeStory.ai, Cupertino, CA, USA 2025 ∼
- Advisor @ CryptoLab, Inc., Seoul, Korea 2025 ∼

- Co-Founder & CTO / Head of Global R&D / Chief Applied Scientist / Senior Fellow @ Gauss Labs, Inc., Palo Alto, CA, USA 2020 ∼ 2023
- Senior Applied Scientist @ Amazon.com, Inc., Vancouver, BC, Canada 2017 ∼ 2020
- Principal Engineer @ Software R&D Center, Samsung Electronics 2016 ∼ 2017
- Principal Engineer @ Strategic Marketing & Sales, Memory Business 2015 ∼ 2016
- Principal Engineer @ DT Team, DRAM Development, Samsung 2012 ∼ 2015
- Senior Engineer @ CAE Team, Memory Business, Samsung, Korea 2005 ∼ 2012
- PhD - Electrical Engineering @ Stanford University, CA, USA 2001 ∼ 2004
- Development Engineer @ Voyan, Santa Clara, CA, USA 2000 ∼ 2001
- MS - Electrical Engineering @ Stanford University, CA, USA 1998 ∼ 1999
- BS - Electrical & Computer Engineering @ Seoul National University 1994 ∼ 1998

Highlight of Career Journey

- BS in Electrical Engineering (EE) @ Seoul National University
- MS & PhD in Electronics Engineering (EE) @ Stanford University
  - *Convex Optimization - Theory, Algorithms & Software*
  - advisor - *Prof. Stephen P. Boyd*
- Principal Engineer @ Samsung Semiconductor, Inc.
  - *AI & Convex Optimization*
  - collaboration with *DRAM/NAND Design/Manufacturing/Test Teams*
- Senior Applied Scientist @ Amazon.com, Inc.
  - *e-Commerce AIs* - anomaly detection, deep RL, and recommender system
  - *Jeff Bezos's project - drove $200M* in sales via Amazon Mobile Shopping App
- *Co-Founder & CTO / Global R&D Head & Chief Applied Scientist* @ Gauss Labs, Inc.
- *Co-Founder & CTO* @ Erudio Bio, Inc.
- *Co-Founder & CEO* @ Erudio Bio Korea, Inc.

Unpacking AI

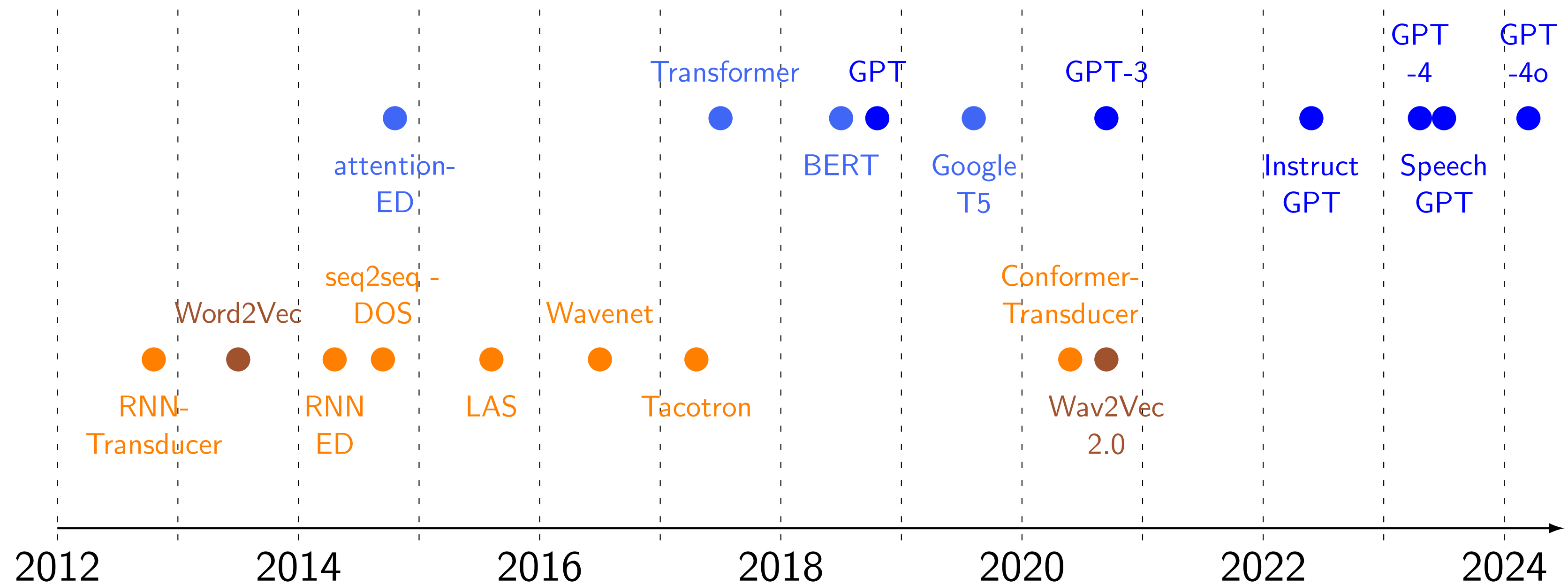
LLM

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-tranining *high performing model* and fine-tuning/transfer learning/domain adaptation
- this *high performing model* learning essential language representation *is* (lanauge) foundation model

– actually, same for other types of learning, $e.g.$, CV

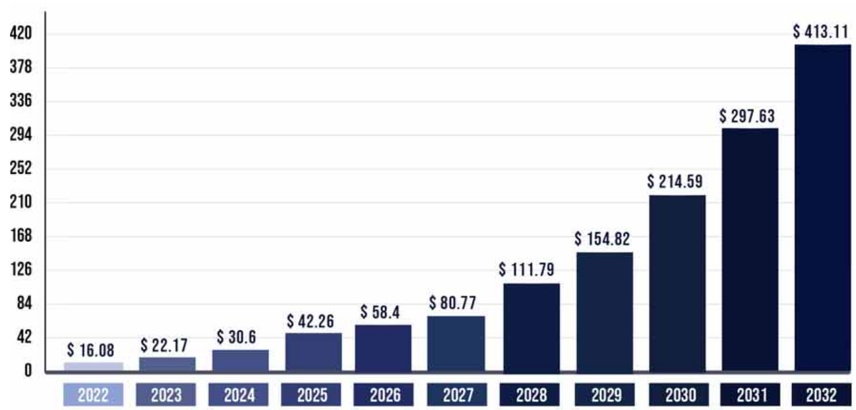
voice recognition	NL generation	summarization	spelling correction	action planning
voice synthesis	translation	topic model	topic model	
	Q&A	NLU		

NLP Market

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

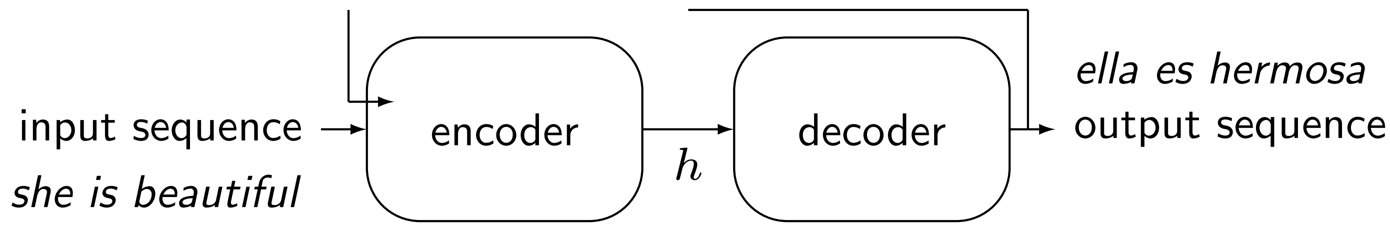
\- source - Precedence Research



Sequence-to-Sequence Models

Sequence-to-sequence (seq2seq) model

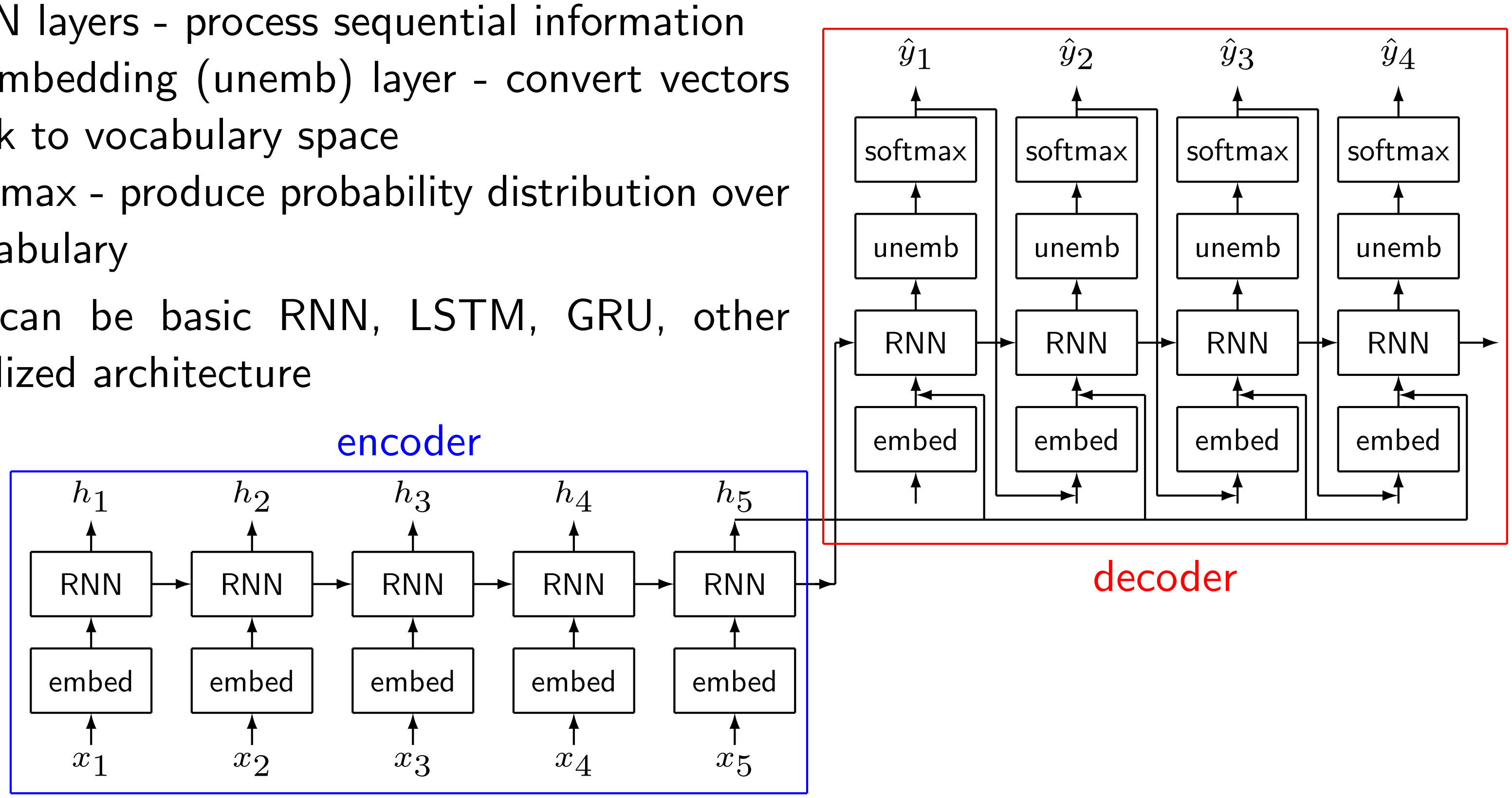
- seq2seq - take sequences as inputs and spit out sequences
- encoder-decoder architecture



  – encoder & decoder can be RNN-type models
  – $h \in \mathbf{R}^n$ - 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

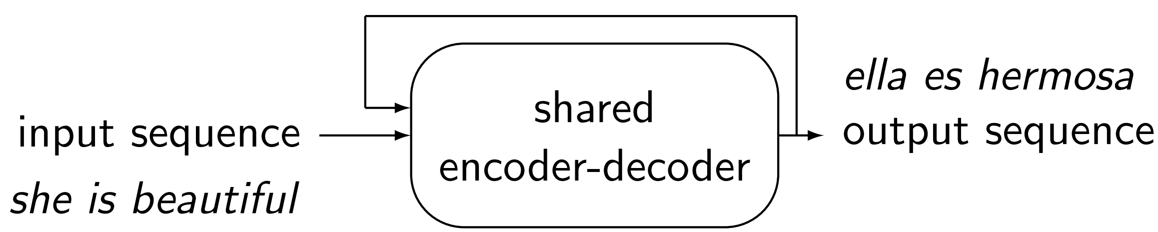
RNN-type encoder-decoder architecture

- components
  - embedding layer - convert input tokens to vector representations
  - RNN layers - process sequential information
  - unembedding (unemb) layer - convert vectors back to vocabulary space
  - softmax - produce probability distribution over vocabulary
- RNN can be basic RNN, LSTM, GRU, other specialized architecture



Shared encoder-decoder model

- single neural network structure can handle both encoding & decoding tasks
  - efficient architecture reducing model complexity
  - allow for better parameter sharing across tasks
- widely used in modern LLMs to process & generate text sequences
  - applications - machine translation, text summarization, question answering
- advantages
  - efficient use of parameters, versatile for multiple NLP tasks



Large Language Models

LLM

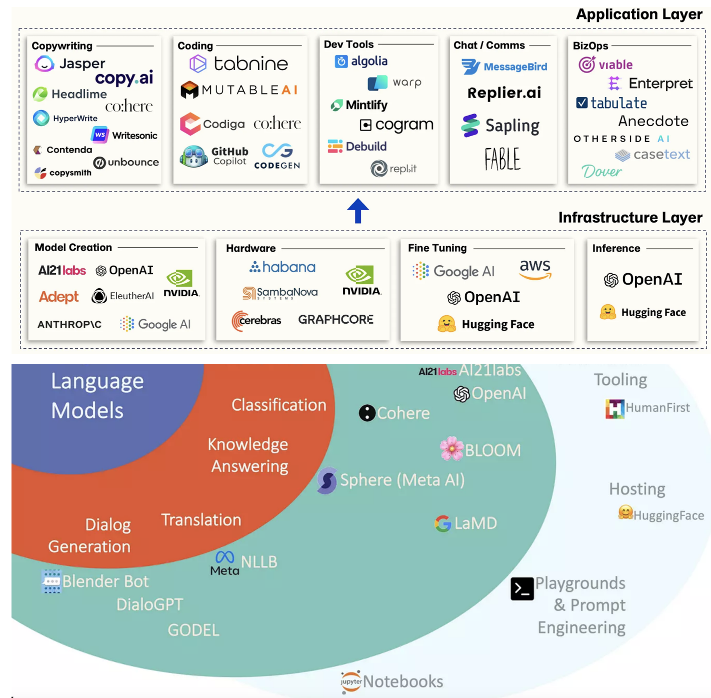
- LLM
  - type of AI aimed for NLP trained on massive corpus of texts & programming code
  - allow learn statistical relationships between words & phrases, $i.e.$, conditional probabilities
  - *amazing performance shocked everyone - unreasonable effectiveness of data (Halevry et al., 2009)*

- applications
  - conversational AI agent / virtual assistant
  - machine translation / text summarization / content creation / sentiment analysis / question answering
  - code generation
  - market research / legal service / insurance policy / triange hiring candidates

  \+ virtually infinite $\#$ of applications

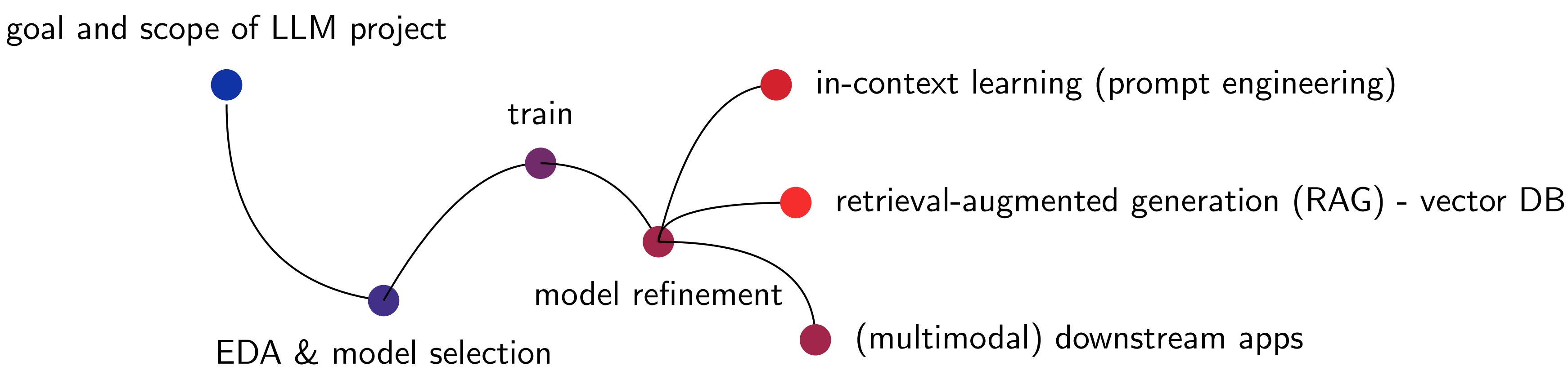
LLMs

- Foundation Models
  - GPT-x/Chat-GPT - OpenAI, Llama-x - Meta, PaLM-x (Bard) - Google
- # parameters
  - generative pre-trained transfomer (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



Transformer

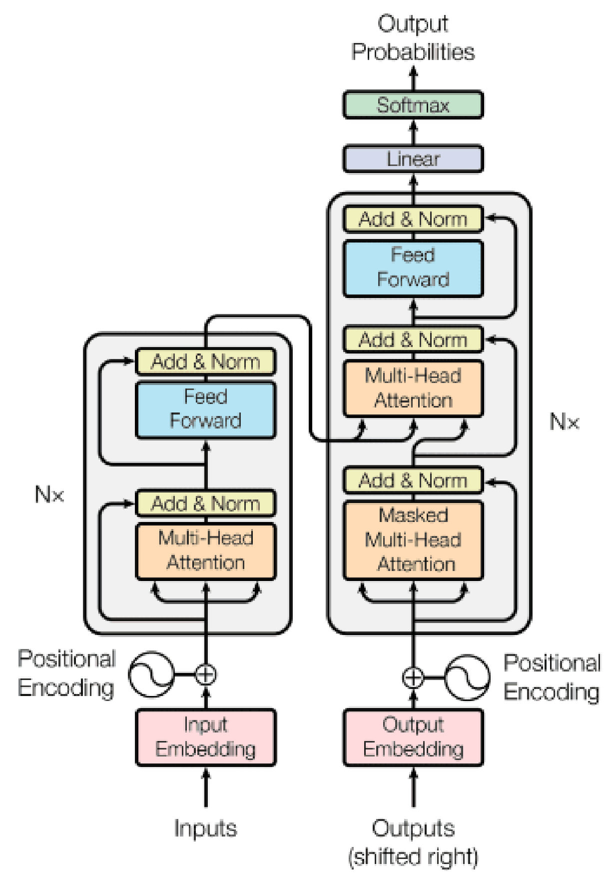
LLM architectural secret (or known) sauce

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)

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

- *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

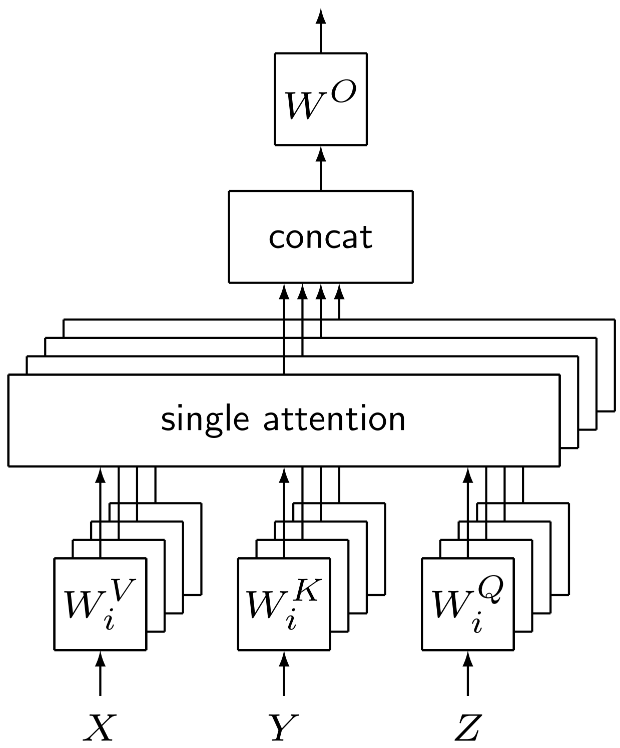
$$X \in \mathbf{R}^{d_e \times m},\ Y \in \mathbf{R}^{d_e \times m},\ Z \in \mathbf{R}^{d_e \times n}$$

  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!*

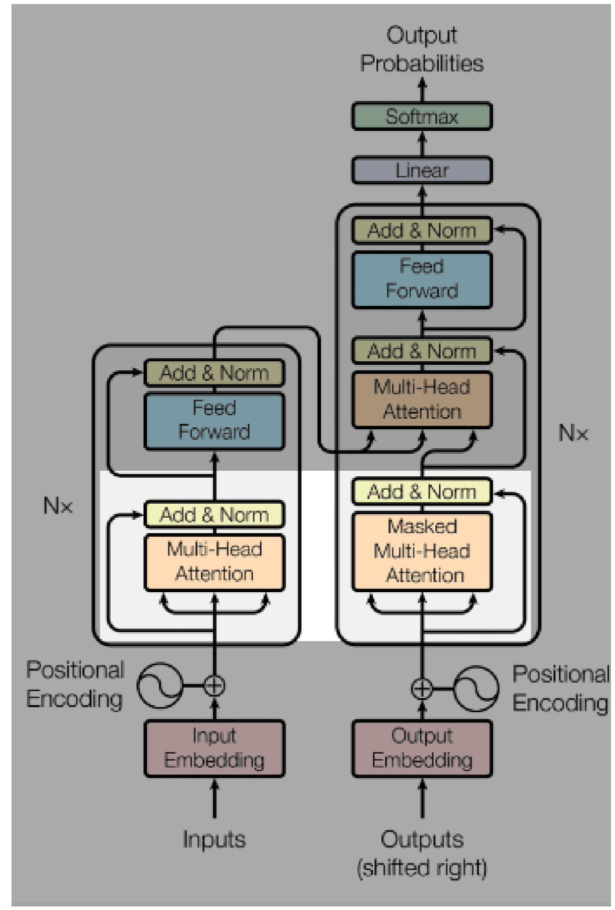
$$W^O \begin{bmatrix} A_1 \\ \vdots \\ A_h \end{bmatrix} \in \mathbf{R}^{d_e \times n},$$

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



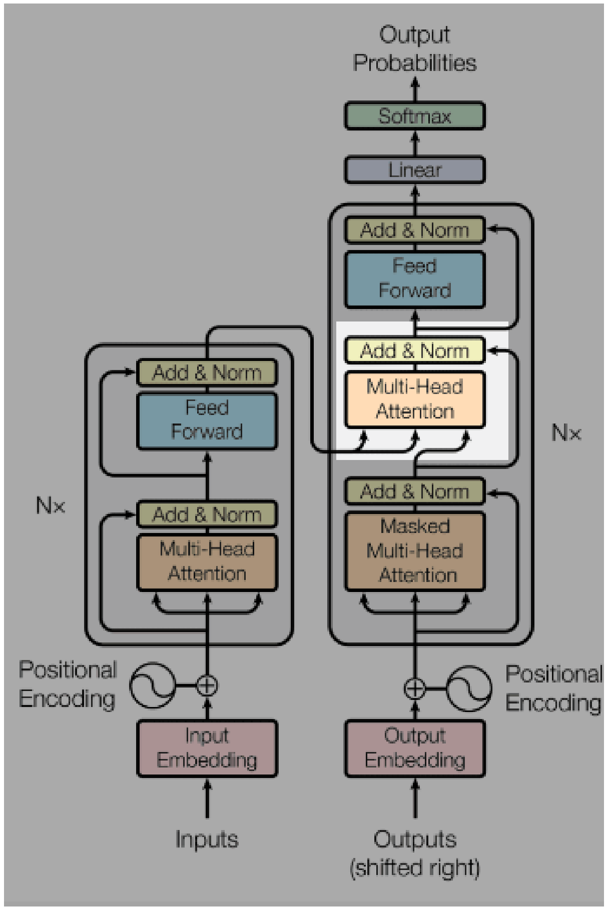
Self attention

- $m = n$
- 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

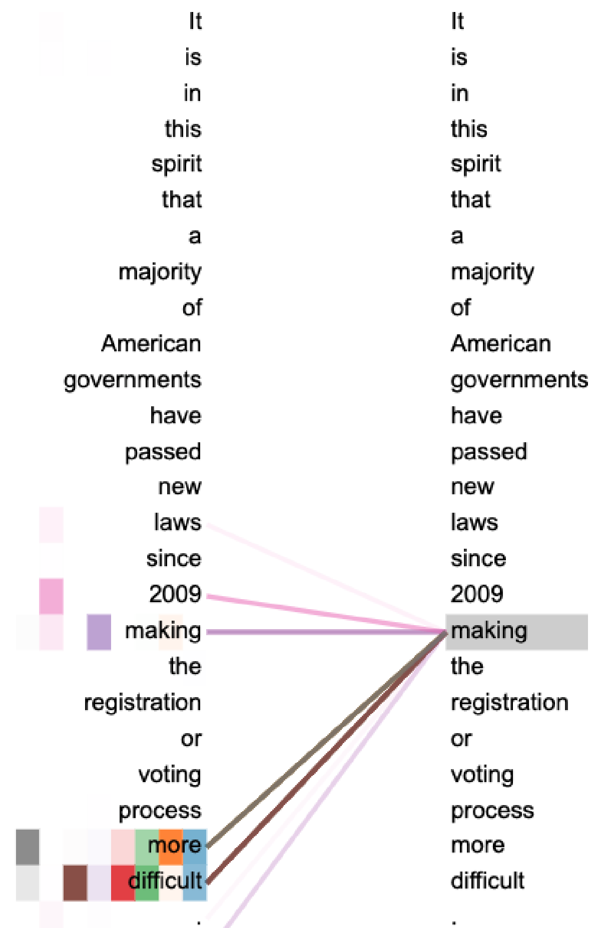
- m: length of input sequence
- n: length of output sequence
- n queries (Q) come from previous decoder layer
- m keys / m values (K, V) come from output of encoder
- every token in output sequence attends to every token in input sequence



Visualization of self attentions

example sentence

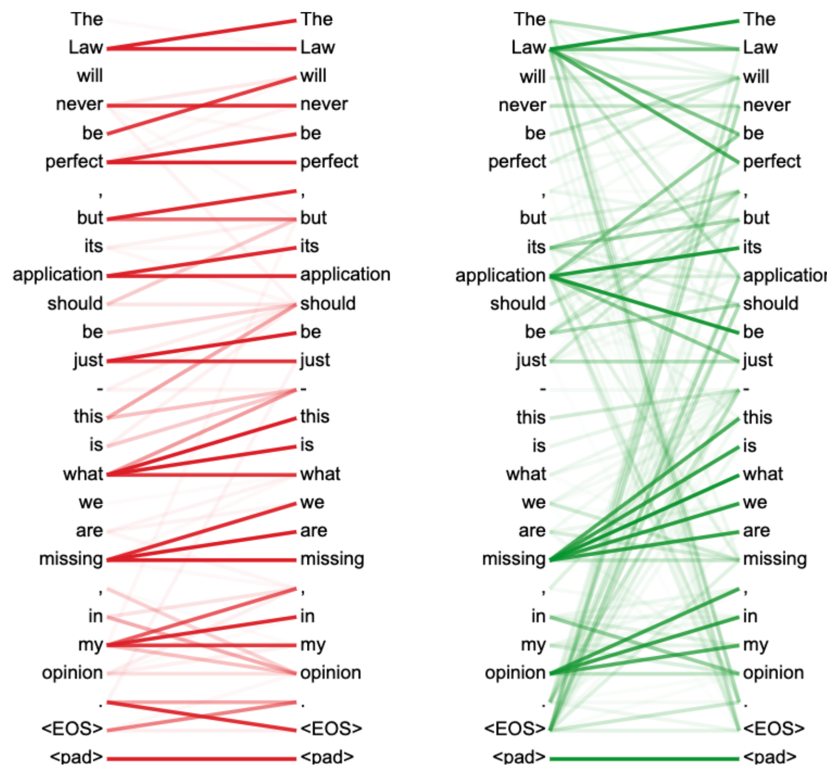
"It is in this spirit that a majority of American governments have passed new laws since 2009 making the registration or voting process more difficult."



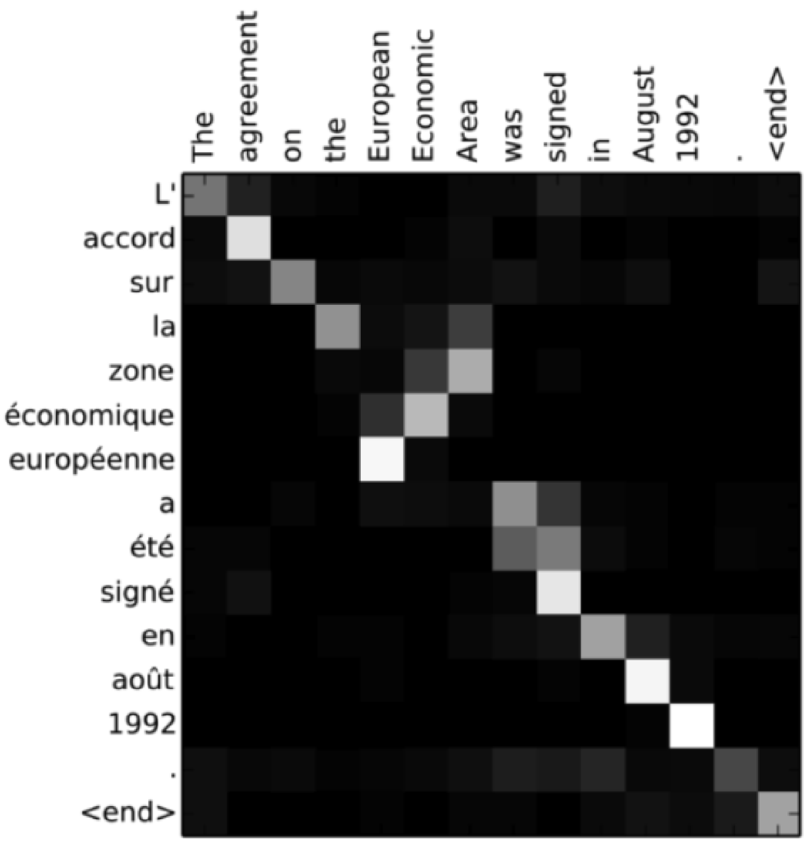
- self attention of encoder (of a layer)
  - right figure
    - show dependencies between "making" and other words
    - different columns of colors represent different heads
  - "making" has strong dependency to "2009", "more", and "difficult"

Visualization of multi-head self attentions

- self attentions of encoder for two heads (of a layer)
  - different heads represent different structures
    $\rightarrow$ 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: $\mathcal{O}(n^2 d)$, 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*

  $\longrightarrow$ *makes NVidia more competitive, hence profitable!*

  $\longrightarrow$ *makes SK Hynix prevail HBM market!*

Variants of Transformer

Bidirectional encoder representations from transformers (BERT)

- Bidirectional Encoder Representations from Transformers [DCLT19]
- pre-train deep bidirectional representations from unlabeled text
- fine-tunable for multiple purposes

input	[cls]	my	dog	is	cute	[sep]	she	likes	playing	[sep]
token embeddings	$E_{\text{[cls]}}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{\text{[sep]}}$	E_{she}	E_{likes}	E_{playing}	$E_{\text{[sep]}}$
	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
segment embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B
	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$	$+$
poistional embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9

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

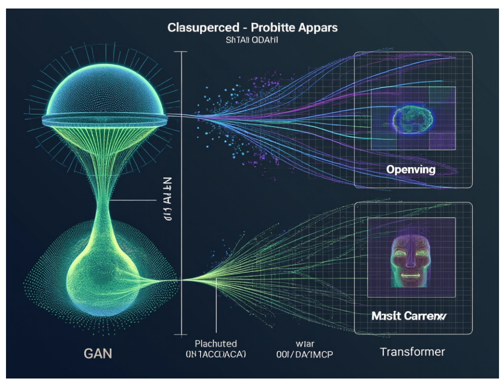
genAI

Definition of genAI

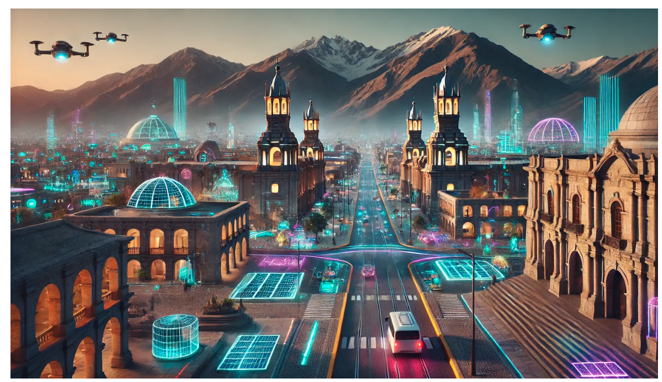
Generative AI

- genAI refers to systems capable of producing new (& original) contents based on patterns learned from training data (representation learning)
  - as opposed to discriminative models for, *e.g.*, classification, prediction & regression
  - here content can be text, images, audio, video, *etc.* - what about smell & taste?
- genAI model examples
  - generative adversarial networks (GANs), variational autoencoders (VAEs), diffusion models, Transformers

by Midjourney



by Grok 2 mini



by Generative AI Lab

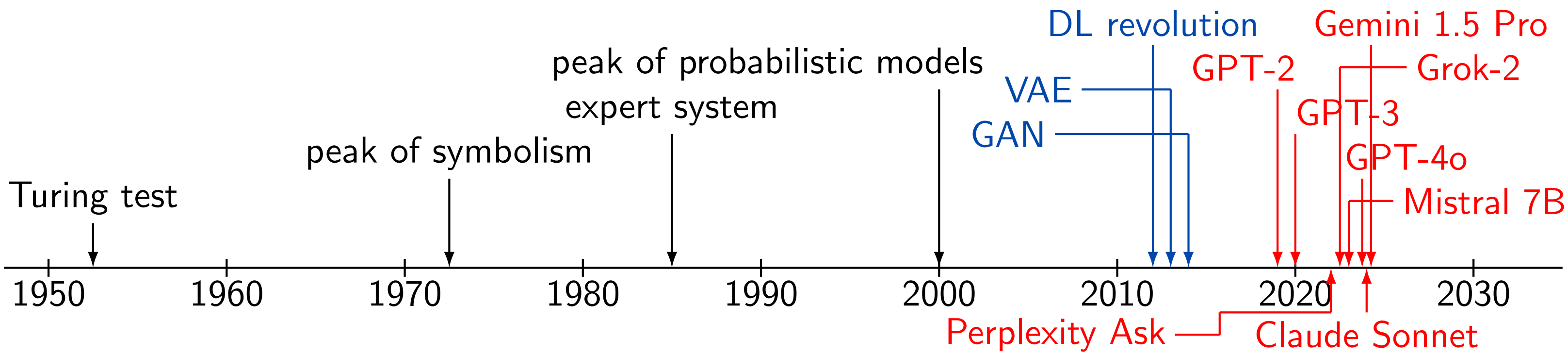
Examples of genAI in action

- text generation
  - Claude, ChatGPT, Mistral, Perplexity, Gemini, Grok
  - conversational agent writing articles, code & even poetry
- image generation
  - DALL-E - creates images based on textual descriptions
  - Stable Diffusion - uses diffusion process to generate high-quality images from text prompts (by denoising random noise)
  - MidJourney - art and visual designs generated through deep learning
- music generation
  - Amper Music - generates unique music compositions
- code generation
  - GitHub Copilot - generates code snippets based on natural language prompts

History of genAI

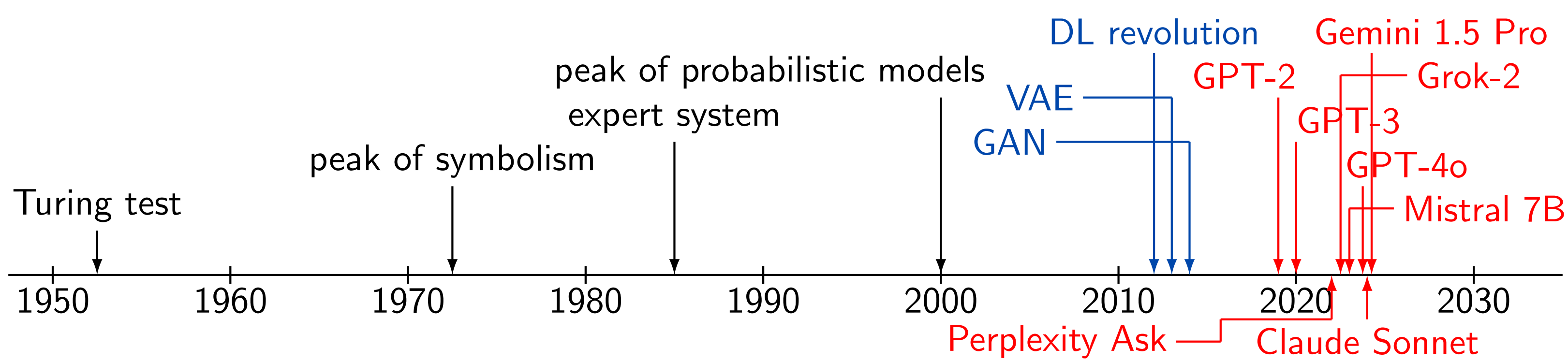
Birth of AI - early foundations & precursor technologies

- 1950s $\sim$ 1970s
  - Alan Turing - concept of *"thinking machine" & Turing test* to evaluate machine intelligence (1950s)
  - *symbolists* (as opposed to connectionists) - early AI focused on symbolic reasoning, logic & problem-solving - Dartmouth Conference in 1956 by *John McCarthy, Marvin Minsky, Allen Newell & Herbert A. Simon*
  - precursor technologies - genetic algorithms (GAs), Markov chains & *hidden Markov models (HMMs)* - laying foundation for generative processes (1970s $\sim$)



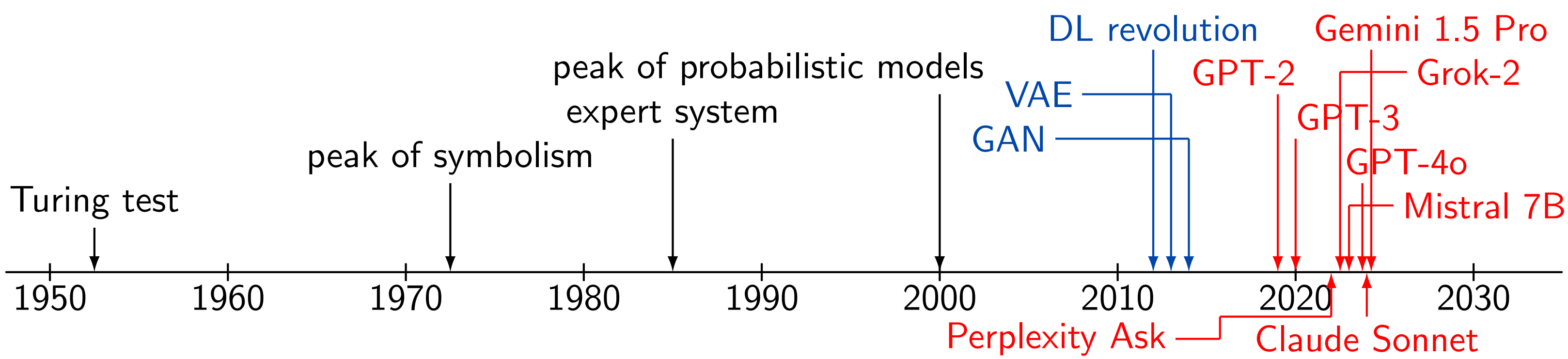
Rule-based systems & probabilistic models

- 1980s $\sim$ early 2000s
  - *expert systems* (1980s) - AI systems designed to mimic human decision-making in specific domains
  - development of neural networks (NN) w/ backpropagation *training multi-layered networks* - setting stage for way more complex generative models
  - *probabilistic models* (including network models, $i.e.$, Bayesian networks) & Markov models - laying groundwork for data generation & pattern prediction



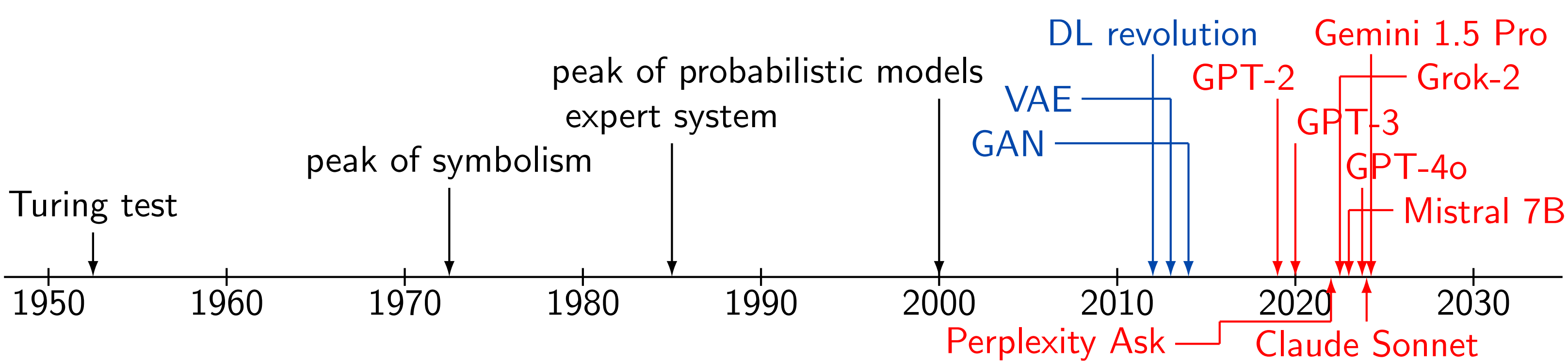
Rise of deep learning & generative models

- 2010s - breakthrough in genAI
  - *deep learning (DL) revolution* - advances in GPU computing and data availability led to the rapid development of deep neural networks.
  - *variational autoencoder (VAE)* (2013) - by Kingma and Welling - learns mappings between input and latent spaces
  - *generative adversarial network (GAN)* (2014) - by Ian Goodfellow - game-changer in generative modeling where two NNs compete each other to create realistic data
    - widely used in image generation & creative tasks



Transformer models & multimodal AI

- late 2010s $\sim$ Present
  - Transformer architecture (2017) - by Vaswani et al.
    - *revolutionized NLP,* $e.g.$, LLM & various genAI models
  - GPT series - generative pre-trained transformer
    - GPT-2 (2019) - generating human-like texts - *marking leap in language models*
    - GPT-3 (2020) - 175B params - set *new standards for LLM*
  - multimodal systems - DALL-E & CLIP (2021) - *linking text and visual data*
  - emergence of diffusion models (2020s) - new approach for generating high-quality images - progressively "denoising" random noise (DALL-E 2 & Stable Diffusion)



Mathy Views on genAI

genAI models

- definition of generative model

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

- *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 plausiable 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) [KW19]

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

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

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

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

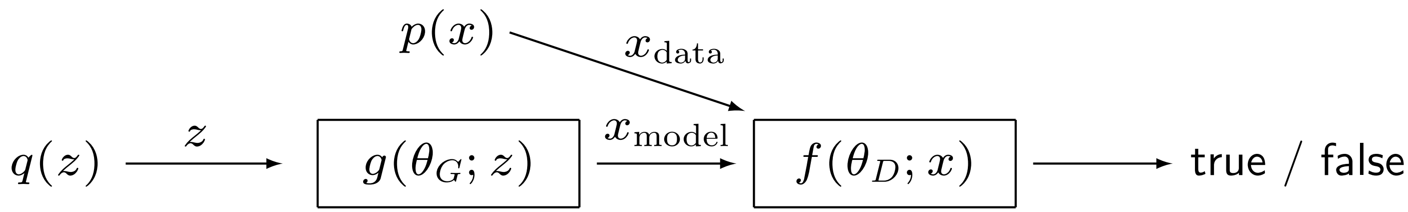
$$\mathcal{L}(\theta,\phi;x) = \mathop{\mathbf{E}}_{z\sim q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

- generative model

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

GAN - early genAI model

- generative adversarial networks (GAN) [GPAM+14]



  - 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

- *maximize 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

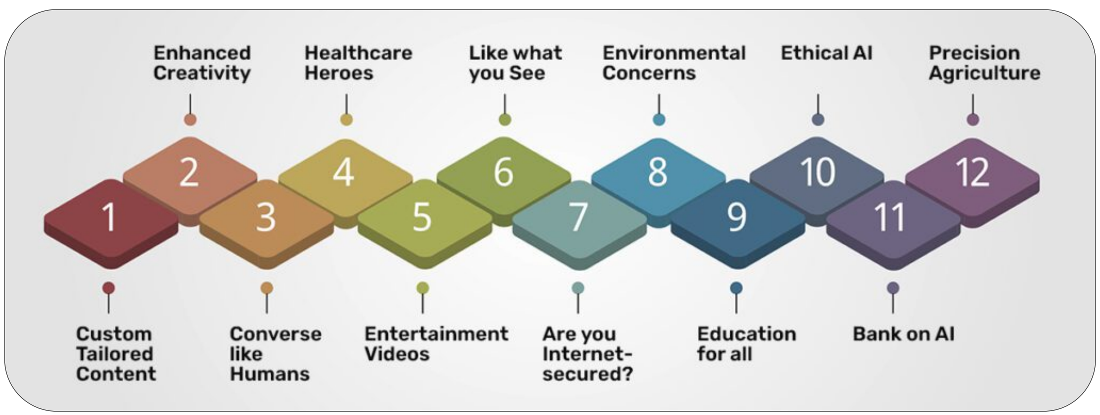
- p_θ represented by (extremely) complicated model
  – $e.g.$, containing multi-head & multi-layer Transformer architecture inside
- model parameters, $e.g.$, for Llama2

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

Current Trend & Future Perspectives

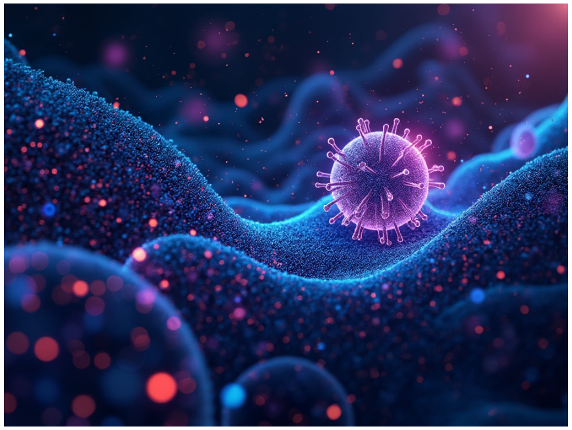
Current trend of genAI

- rapid advancement in language models & multimodal AI capabilities
- rise of AI-assisted creativity & productivity tools
- growing adoption across industries
  - creative industries - design, entertainment, marketing, software development
  - life sciences - healthcare, medical, biotech
- infrastructure & accessibility, *e.g.*, Hugging Face democratizes AI development
- integration with cloud platforms & enterprise-level tools
- increased focus on AI ethics & responsible development



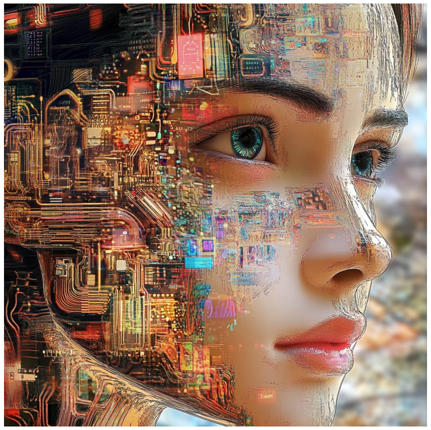
Industry & business impacts

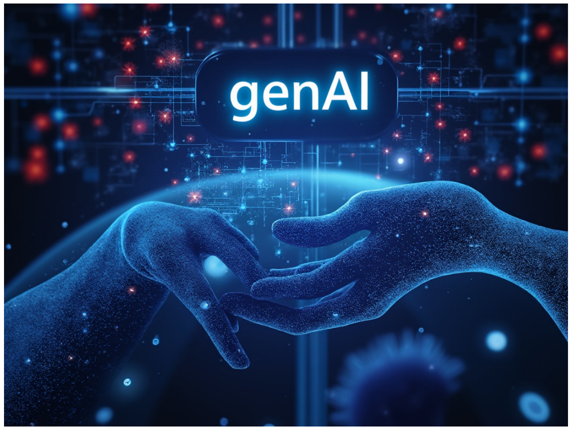
- how genAI is transforming industries
  - creative industries - content creation - advertising, gaming, film
  - life science - enhance research, drug discovery & personalized treatments
  - finance - automating document generation, risk modeling & fraud detection
  - manufacturing & Design - rapid prototyping, 3D modeling & optimization
  - business operations - automate routine tasks to boost productivity



Future perspectives of genAI

- hyper-personalization - highly personalized content for individual users - music, products & services
- AI ethics & governance - concerns over deepfakes, misinformation & bias
- interdisciplinary synergies - integration with other fields such as quantum computing, neuroscience & robotics
- human-AI collaboration - augment human creativity rather than replace it
- energy efficiency - have to figure out how to dramatically reduce power consumption





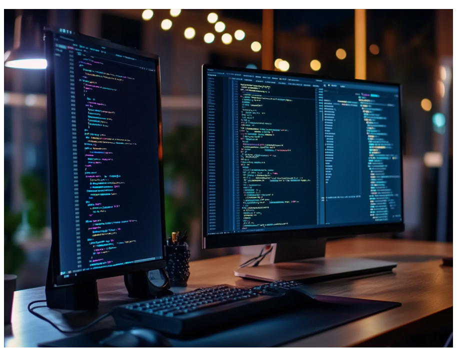
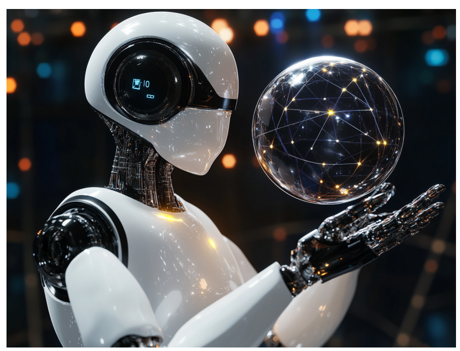
AI Products

AI product development - trend and characteristics

- *rapid pace* of innovation - new AI models & products being released at unprecedented rate, improvements coming in weeks or months (rather than years)
- *LLMs dominating* - models like GPT-4 & Claude pushing boundaries in NLP & genAI
- *multimodal AI* gaining traction - models processing & generating text, images & even video becoming more common, *e.g.*, Grok, GPT-4, Gemini w/ vision capabilities
- *open-source* AI movement - growing trend of open-source AI models and tools, challenging dominance of proprietary systems
- *AI integration in everyday products* - from smartphones to home appliances, AI being integrated into wide array of consumer products

AI product development - trend and characteristics

- *ethical AI & regulatory focus* - increased attention on ethical implications of AI & calls for regulation of AI development and deployment
- AI in enterprise - businesses across industries rapidly adopting AI for various applications
- *specialized AI models* - development of AI models tailored for specific industries or tasks, $e.g.$, healthcare, biotech, financial analysis
- AI-assisted *coding and development* - help software developers write code more efficiently & tools becoming increasingly sophisticated
- *concerns about AI safety & existential risk* - growing debate about potential short & long-term risks of advanced AI


LLM products

- OpenAI - ChatGPT 4o, GPT-4 Turbo Canvas
- Anthropic - Claude 3.5 Sonnet (with Artifacts), Claude 3 Opus, Claude 3 Haiku
- Mistral AI - Mistral 7B, Mistral Large 2, Mistral Small `xx.xx`, Mistral Nemo (12B)
- Google - Gemini (w/ 1.5 Flash), Gemini Advanced (w/ 1.5 Pro)
- X - Grok [mini] [w/ Fun Mode]
- Perplexity AI - Perplexity [Pro] - combines GPT-4, Claude 3.5, and Llama 3
- Liquid AI - Liquid-40B, Liquid-3B (running on small devices)

flying cats generated by Grok, ChatGPT 4o & Gemini

Comparison of LLMs & LLM products

model	developer	training data	# params	strength	weakness
GPT-4	OpenAI	web & books	170B	advanced reasoning & multimodal capabilities	high computational resources
LLaMA-2	Meta	public info & research articles	7∼70B	open access & good performance for different sizes	not powerful for complex tasks
Claude	Anthropic	mix of high-quality datasets	not disclosed	safety-first approach avoiding harmful responses	limited in publicly available details
PaLM 2	Google	multilingual text corpus	540B	high multilingual comprehension supporting various downstream apps	significant resources & not versatile in some contexts

Comparison of LLMs & LLM products

model	developer	training data	# params	strength	weakness
BLOOM	BigScience Community	diverse multilingual corpus	176B	open & support multiple languages	resource-intensive & lower performance
Mistral[1]	Mistral AI	public web data	7∼13B	lower parameter count	limited scalability for specialized apps
Liquid Foundation Model (LFM)	Liquid AI	adaptive datasets	adaptive & dynamic parameters	modular & support more specialized fine-tuning for niche use-cases & adaptable in deployment	complexity in design and implementation

Multimodal genAI products

- DALL-E by OpenAI
  - *generate unique and detailed images based on textual descriptions*
  - understanding context and relationships between words
- Midjourney by Midjourney
  - let people *create imaginative artistic images*
  - can interactively guide the generative process, providing high-level directions

Multimodal genAI products

- Dream Studio by Stability AI
  - *analyze patterns in music data & generates novel compositions*
  - musicians can explore new ideas and enhance their *creative* processes
- Runway by Runway AI
  - *realistic images, manipulate photos, create 3D models & automate filmmaking*

Rise of co-pilot products

- definition - AI-powered tools designed to enhance human productivity across multiple domains including document creation, presentations & coding
- benefits
  - *efficiency* - automate repetitive tasks allowing users to focus on high-value activities
  - *error reduction* - minimize mistakes common in manual work
  - *creativity* - suggestions and prompts help users explore new ideas and approaches
  - *integration* with major productivity suites - Microsoft 365, Google Workspace
- popular products
  - GitHub Copilot, Microsoft 365 Copilot, Grammarly AI, Visual Studio Code Extensions



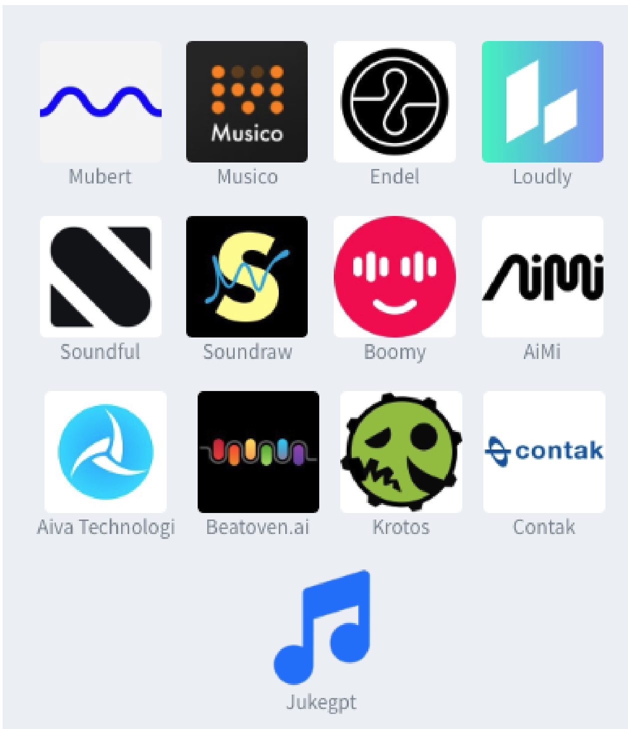
Future of co-pilot products

- potential advancements
  - wider adoption across industries and professions
  - *real-time fully automated collaboration*, *predictive content generation*, personalization
- impact on work environments & creative processes
  - *collaborative human-AI relationships* with augmented reality
  - unprecedented levels of problem-solving due to *augmented cognitive abilities*
- challenges & considerations
  - *ethical concerns around data privacy & AI decision-making*
  - potential impact on *human skills & job markets*

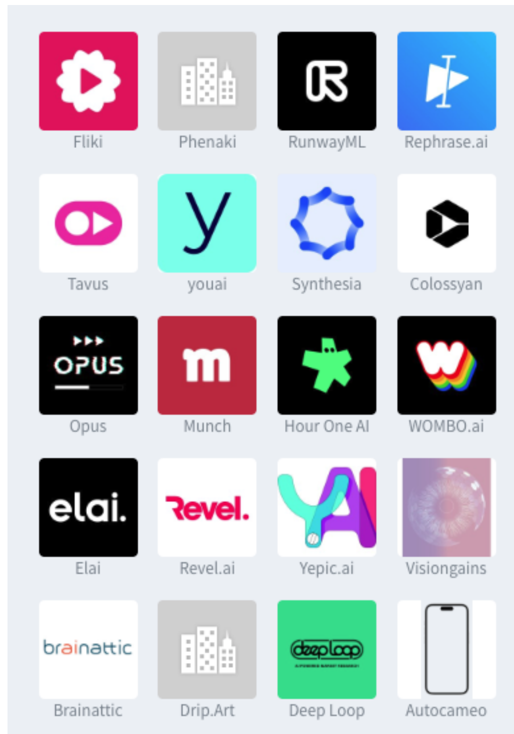


Other AI products - audio/video/text

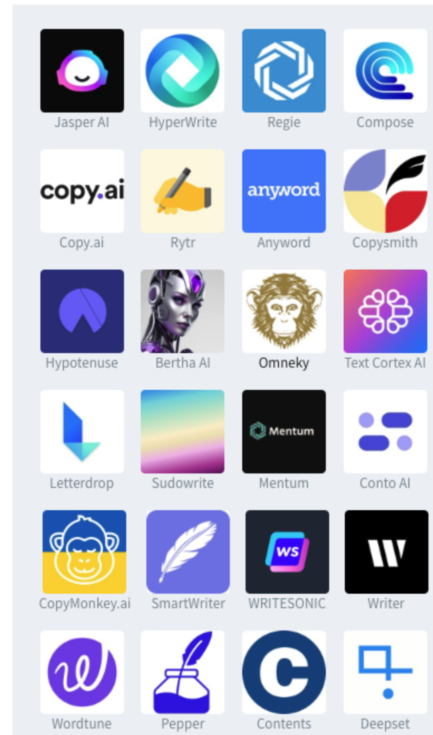
audio



vidio

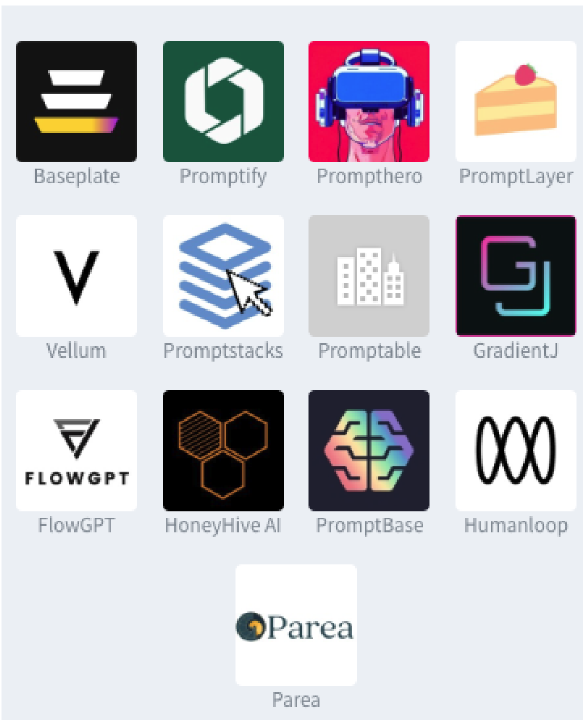


text

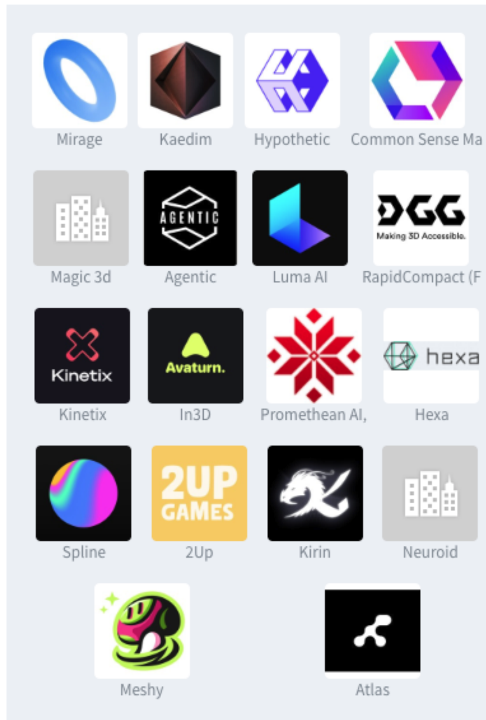


Other AI products - LLM/gaming/design/coding

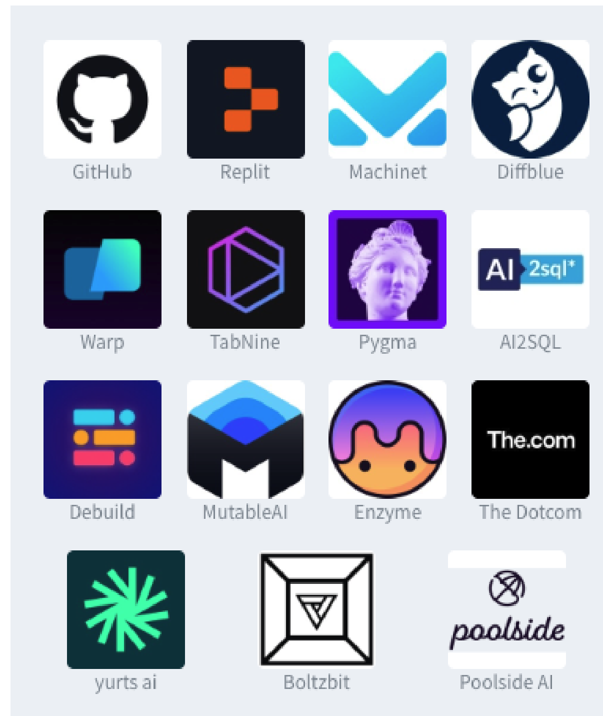
LLM



gaming & design



coding



AI Market & Values

AI market

- PwC, one of "big four" accounting firms, believes
  - *AI can add $15.7 trillion to the global economy by 2030*

Cloud stacks

- SaaS dominates cloud stack - account for 40% of total cloud stack market with estimated TAM of $260B
- IaaS and PaaS significant players
- semi-cloud's niche presence

cloud stack	companies	estimated TAM	% total in stack
SaaS apps	Salesforce, Adobe	$260B	40%
PaaS	Confluent, snowflake	$140B	22%
IaaS	AWS, Azure, GCP	$200B	30%
cloud semis	AMD, Intel	$50B	8%

AI stacks

- AI investment landscape - AI sector witnessing significant capital inflow with total funding of approximately $29 billion across various segments
- models lead pack - AI models, particularly those developed by OpenAI and Anthropic, attracted lion's share of investments, accounting for 60% of total funding
- diverse growth - while models dominate funding, other segments like apps, AI cloud, and AI semis also experiencing substantial growth, indicating broadening AI ecosystem

AI stack	companies	total funding	% total in stack
apps	character.io, replit	~$5B	17%
models	openAI, ANTHROP\C	~$17B	60%
AIops	Hugging Face, Weights & Biases	~$1B	4%
AI cloud	databricks, Lambda	~$4B	13%
AI semis	cerebras, SambaNova	~$2B	6%

AI model companies

- AI model companies - competing for which AI model companies will dominate 2020s
- venture funding surge - private AI model companies raised approximately $17B since 2020, indicating strong investor confidence
- growing open-source presence - becoming increasingly prevalent, adding competition and innovation to AI landscape
- key players - notable companies in AI model space include Adept, OpenAI, Anthropic, Imbue, Inflection, Cohere, and Aleph Alpha
- outcome uncertain - future success is still to be determined, reflecting dynamic and evolving nature of AI industry

AI advancing much faster

- rapid AI advancement - general AI projected to progress from basic content generation to superhuman reasoning in only 5 years
- significantly outpacing 15-year timeline for fully autonomous vehicles

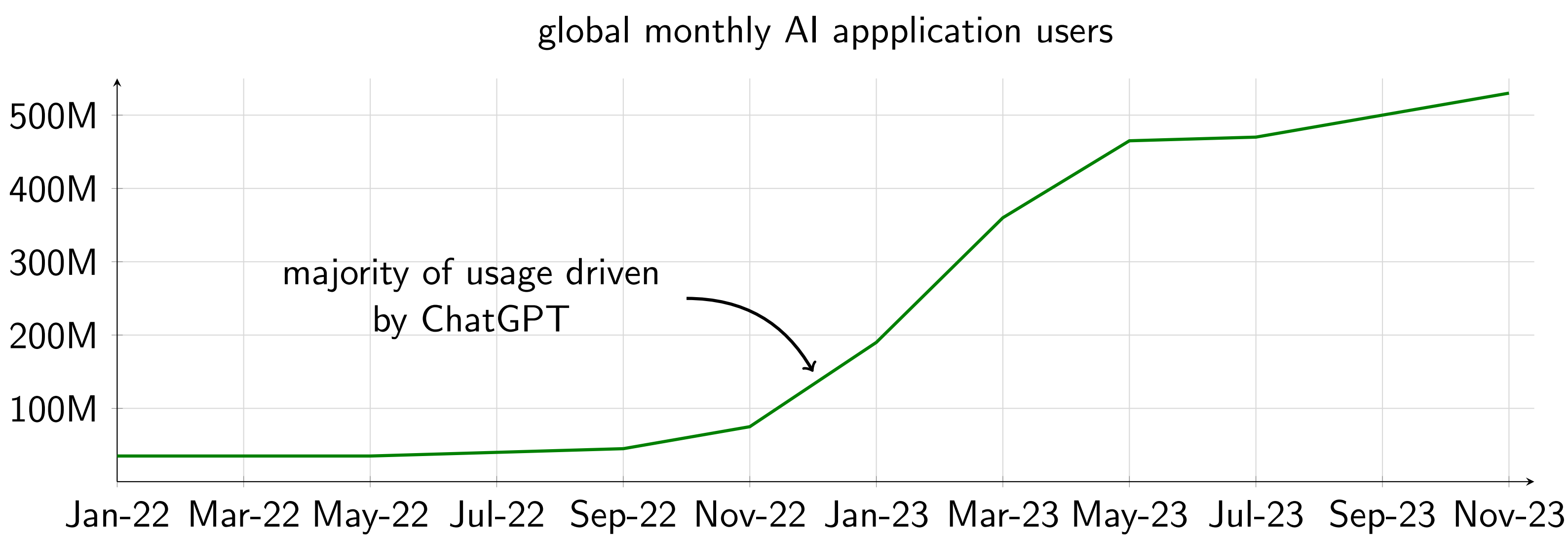
autonomy level	autonomous vehicles	genAI
L5	fullly autonomous	superhuman reasoning & perception
L4	highly autonomous	AI autopilot for complex tasks
L3	self-driving with light intervention	AI co-pilot for skilled labor
L2	Tesla autopilot	supporting humans with basic tasks
L1	cruise control	generating basic content

15 yrs

5 yrs

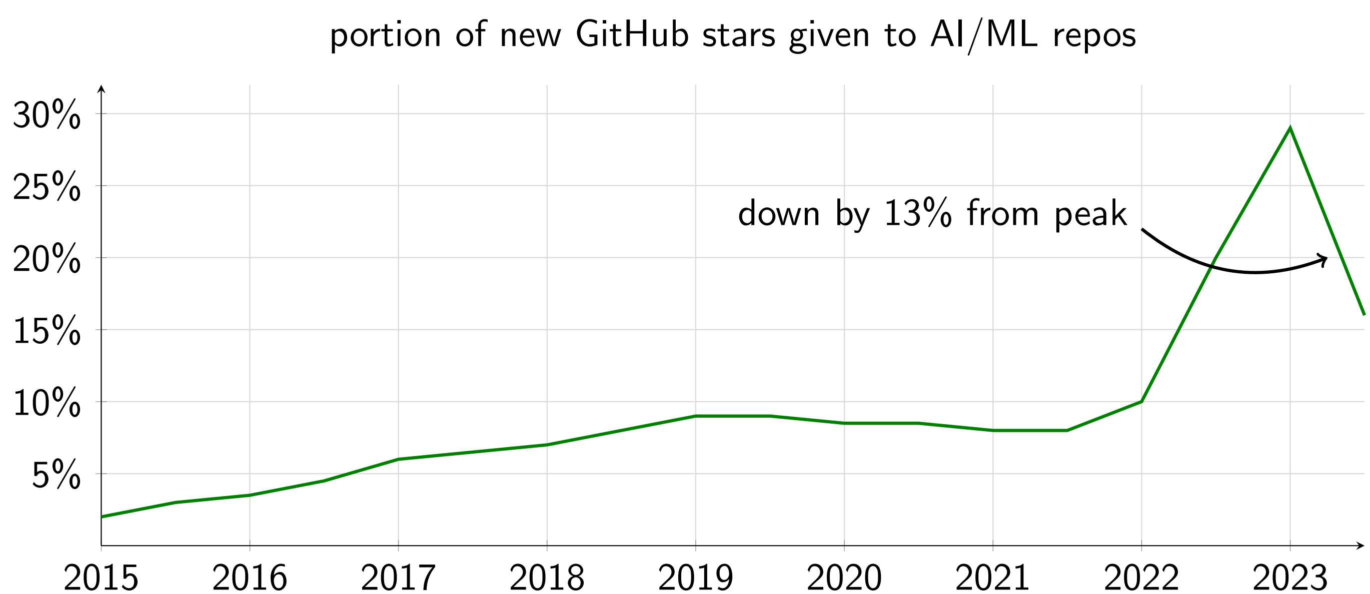
AI interest of users

- AI adoption approaching saturation - initial wave may be nearing saturation
- future growth might come from deeper integration into professional workflows & specialized applications
- potential for market diversification - ChatGPT drove majority of early growth, but now we have other LLMs - Claude, Mistral, Gemini, Grok, Perplexity



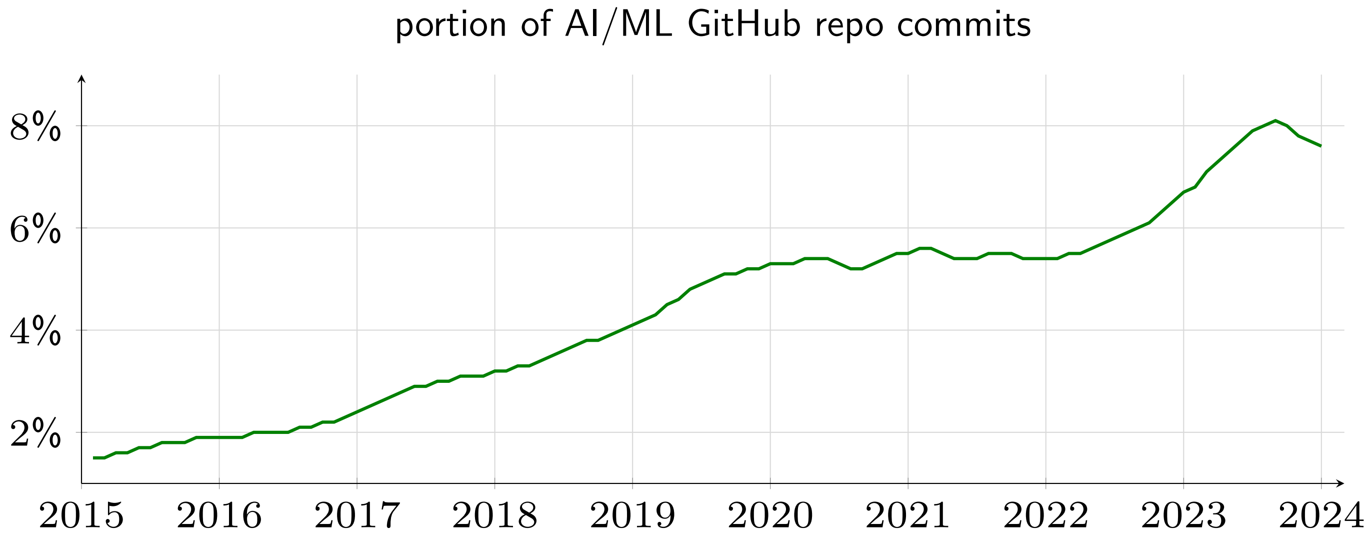
AI interest of developers

- rising popularity - portion of new GitHub stars given to AI/ML repositories steadily increased from 2015 to 2022
- excitement waning & washing out AI "tourists" - decline of 13% from peak in 2022
- could indicate potential factors such as market saturation, economic conditions, or shifts in developer preferences



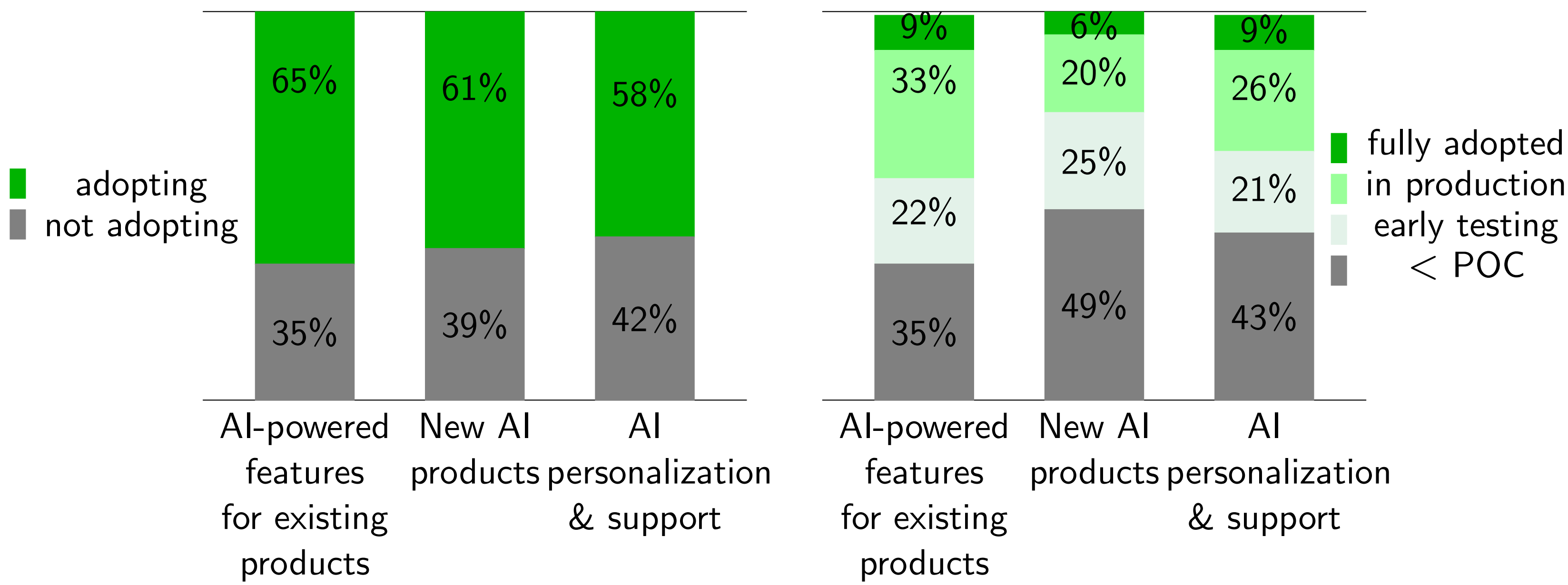
Developers' contribution to software packages

- steep acceleration from 2022 to 2024 correlates with explosion of LLMs & genAI
- suggesting transformative shift in AI landscape beyond gradual growth
- AI/ML still represents relatively small portion (less than 10%)
- indicating significant room for growth and mainstream adoption across various software domains



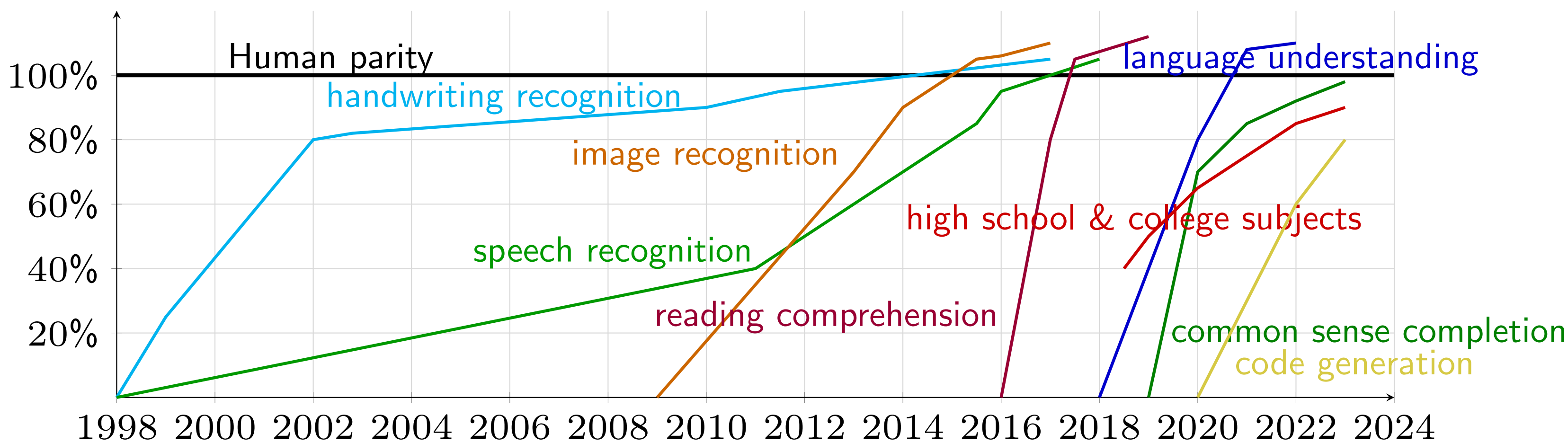
Enterprises adoptiong AI

- more than 60% of enterprises planning to adopt AI
- full adoption rate is less than 10% - will take long time



AI getting better and faster

- steep upward slopes of AI capabilities highlight accelerating pace of AI development
  - period of exponential growth with AI potentially mastering new skills and surpassing human capabilities at ever-increasing rate
- closing gap to human parity - some capabilities approaching or arguably reached human parity, while others having still way to go
  - achieving truly human-like capabilities in broad range remains a challenge



AI delivers game-changing values

- time developers save using GitHub Copilot - *55%*
  - *10M+* cumulative downloads as of 2024 & *1.3M* paid subscribers - *30%* Q2Q increase
  - improves developer productivity by *30%+*
- reduction in human-answered customer support requests - *45%*
  - cost per support interaction - *95%* save / $2.58 (human) vs $0.13 (AI)
  - median response time - *44 min* faster / 45 min (human) vs 1 min (AI)
  - median customer satisfaction - *14%* higher / 55% (human) vs 69% (AI)
- time saved from editing video in runway - *90%*
- AI chat rated higher quality compared to physician responses - *79%*

Selected References & Sources

Selected references & sources

- Robert H. Kane "Quest for Meaning: Values, Ethics, and the Modern Experience" 2013
- Michael J. Sandel "Justice: What's the Right Thing to Do?" 2009
- Daniel Kahneman "Thinking, Fast and Slow" 2011
- Yuval Noah Harari "Sapiens: A Brief History of Humankind" 2014
- M. Shanahan "Talking About Large Language Models" 2022
- A.Y. Halevry, P. Norvig, and F. Pereira "Unreasonable Effectiveness of Data" 2009
- A. Vaswani, et al. "Attention is all you need" @ NeurIPS 2017
- S. Yin, et. al. "A Survey on Multimodal LLMs" 2023
- Chris Miller "Chip War: The Fight for the World's Most Critical Technology" 2022

- CEOs, CTOs, CFOs, COOs, CMOs & CCOs @ startup companies in Silicon Valley
- VCs on Sand Hill Road - Palo Alto, Menlo Park, Woodside in California, USA

References

References

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[GPAM+14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[KW19] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, 12(4):307–392, 2019.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of 31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

Thank You