Introduction to Multimodal Deep Learning

Multimodal Generative AI Theories and Applications

Lecture 1

Jin-Hwa Kim and Sangdoo Yun

Ref. Gardner (1985) Ch. 6

NAVER AI LAB

Table of contents

Class introduction

Course objectives

- Gain a deep understanding of multimodal generative AI theories and applications.
- Explore various modalities, including vision, language, audio, and speech, in multimodal deep learning.
- Stay updated with emerging topics in neural graphics, specifically NeRFs and 3DGS.

Co-lecturers

Sangdoo Yun

Research Director at NAVER AI Lab
Guest Assistant Professor at SNU AIIS
Working on various ML models
for real-world applications.

Jin-Hwa Kim

Leader at NAVER AI Lab
Guest Associate Professor at SNU AIIS
Working on multimodal generation models
focusing on neural 3D approaches.

General information

- Co-lectured by Sangdoo Yun & Jin-Hwa Kim
- Friday 1 PM Sep. 5 to Dec. 19 (schedule table in the next slide)
- Office hour: Friday 10 AM 942-208 (AIIS)
  - Sangdoo Yun (sangdoo.yun@navercorp.com)
  - Jin-Hwa Kim (jnhwkim@snu.ac.kr)
  - *Reservation only*

Schedule

Week	Date	Topic	Note
1	Sep 5	Introduction	Jin-Hwa Kim
2	Sep 12	Multimodal Representation Learning	Jin-Hwa Kim
3	Sep 19	Multimodal Foundation Models I	Sangdoo Yun
4	Sep 26	Multimodal Foundation Models II	Sangdoo Yun
5	Oct 3	National Foundation Day (No Class)	-
6	Oct 10	Multimodal Generation I	*Guest: Junho Kim*
7	Oct 17	Multimodal Foundation Models III	Sangdoo Yun
8	Oct 24	Trustworthy AI (Recorded Video)	Sangdoo Yun
9	Oct 31	**Midterm Exam**	-
10	Nov 7	Neural Graphics I	Jin-Hwa Kim
11	Nov 14	Multimodal Generation II	*Guest: Gayoung Lee*
12	Nov 21	Neural Graphics II	Jin-Hwa Kim
13	Nov 28	Neural Graphics III	Jin-Hwa Kim
14	Dec 5	Speech Generation	Guest: Eunwoo Song
15	Dec 12	Human-Computer Interaction	Guest: Youngho Kim
16	Dec 19	**Final Exam**	-

*ICCV 2025 will be held from Oct 19 to 23, and NeurIPS 2025 will be held from Nov 30 to Dec 5.

Grade

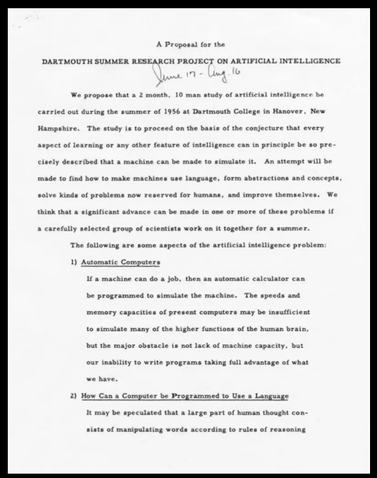
- Attendance *(10%)*
- Attitude *(10%)*
  - Asking a thoughtful question (as determined by the instructor or invited speaker) earns *1 (good)* or *2 (excellent)* points.
  - Giving a short presentation (12 minutes + 3 minutes Q&A) on a topic related to multimodal generative AI earns *1–5 points*. Up to *12* presentation opportunities are available in advance of the scheduled lectures. First come, first served. Students can also earn points by asking questions during these presentations.
  - *The base score is 5 points* and may be deducted at the lecturer's discretion.
- Midterm and final exams *(mid 40% + final 40% = 80%)*

The birth of AI field

The summer of 1956 at Dartmouth

- In the summer of 1956, young scholars gathered at Dartmouth College in Hanover, New Hampshire, the founding event of *artificial intelligence as a field*.
- *"The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it."* (McCorduck* 1979, p.93)

Marvin Minsky, Claude Shannon, Ray Solomonoff and others at the Dartmouth (Photo: Margaret Minsky)



A Proposal for the

DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE

June 17 - Aug 16

We propose that a 2 month, 10 man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

The following are some aspects of the artificial intelligence problem:

1) Automatic Computers

If a machine can do a job, then an automatic calculator can be programmed to simulate the machine. The speeds and memory capacities of present computers may be insufficient to simulate many of the higher functions of the human brain, but the major obstacle is not lack of machine capacity, but our inability to write programs taking full advantage of what we have.

2) How Can a Computer be Programmed to Use a Language

It may be speculated that a large part of human thought consists of manipulating words according to rules of reasoning

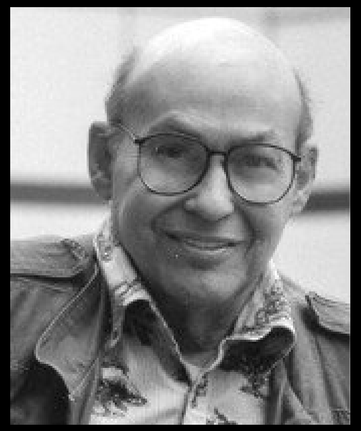
*Pamela McCorduck, a historian of artificial intelligence

Early AI contributors

- John McCarthy (1927-2011)
  - Then, an assistant professor of mathematics at Dartmouth and eventually
  - the founder and first director of the AI Labs at both MIT (1957) and Stanford (1963)
  - the major coiner of the term *artificial intelligence (AI)*,
  - developed the programming language family *LISP* and invented *garbage collection*.
  - 1971 Turing Awardee.

Early AI contributors

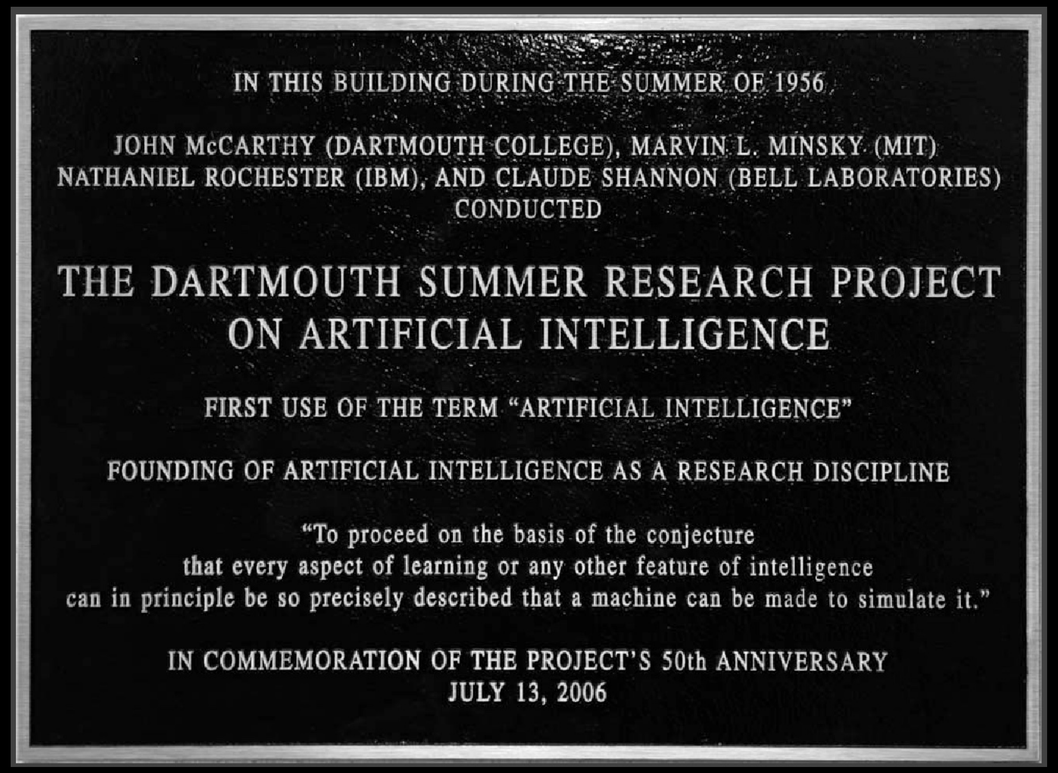
- Marvin Minsky (1927-2016)
  - Then, he was a junior fellow in mathematics and neurology at Harvard and eventually
  - a co-founder of MIT's AI Lab, and the director of the AI Lab at MIT
  - 1969 Turing Awardee.
  - The co-authored book *Perceptrons* attacked *Frank Rosenblatt*'s work, becoming the seminal work in the analysis of artificial neural networks.
  - It was crucial in discouraging neural network research in the 1970s and contributing to the so-called "*AI winter*."



Meaning of Dartmouth

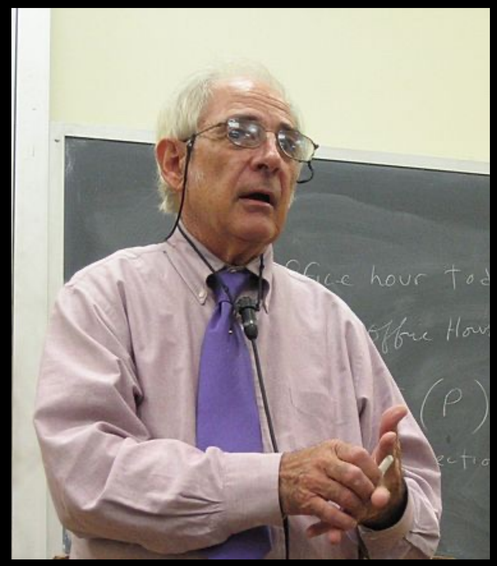
- Despite of limited success, that was symbolic advance from an older generation — Norbert Wiener, John von Neumann, Warren McCulloch, and *Alan Turing* — where in doubt in the development of AI, more than electronic computers.

- *"While no single event can lay claim to signaling the birth of all of cognitive science, the workshop at Dartmouth is the chief contender within the field of AI."*



Conflicting ideas

- Strong vs. weak AI
  - "The appropriately programmed computer really *is* mind"? (John Searle, 1980)
- Generalists vs. experts
- *"The dramatic promise of a thinking machine, the battles about the scientific status of AI have been particularly vehement."*



Photograph of John Searle by Matthew Breindel

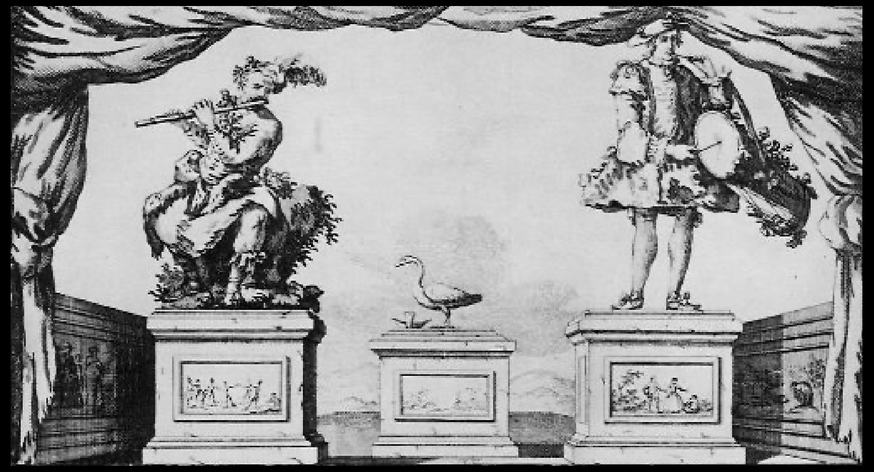
Dream of AI 🇫🇷

- René Descartes (1596-1650) was interested in the automata simulating human body, while skeptical about simulating the mind.
- L'Homme Machine (La Mettrie, 1747)
  - "the human body is a machine that winds up its own springs" (McCorduck, 1979)

René Descartes and Julien Offray de La Mettrie

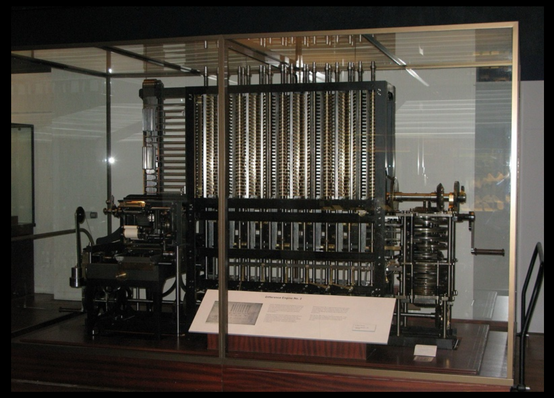
The father of simulation 🇫🇷

- Jacques de Vaucanson (1709–1782) was an automata builder who thrilled Europe with mechanical flute players, ducks, and pipe and taber players.
- Vaucanson *"was concerned to formulate and validate — in the most precise and formal language available to him — a theory of the German flute player"* (Fryer and Marchall, 1979)



British ambitions 🇬🇧

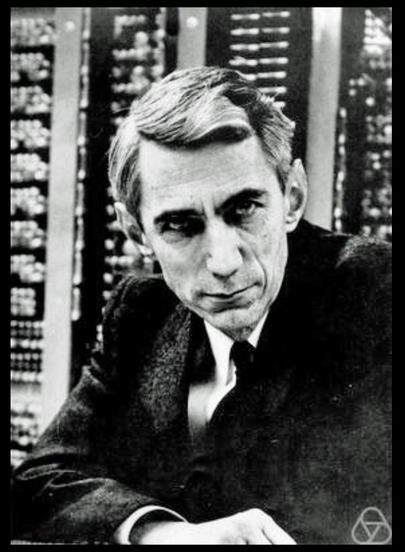
- Charles Babbage (1791–1871)
  - Automatic table calculator for any arithmetic problem, *"difference machine"*
- George Boole (1815–1864)
  - Invent basic laws of thought on the principles of *logic*, *"mental algebra"*
  - Used a set of *symbols* (*e.g.*, a, b, x, y) to stand for the components of thought
  - *"a successful attempt to express logical propositions by symbols, the laws of whose combinations should be founded upon the laws of the mental processes which they represent, would, so far, be a step toward the philosophical language"* (Boole, 1847, p.5)



The London Science Museum's difference engine was built from Babbage's design.

The father of information theory 🇺🇸

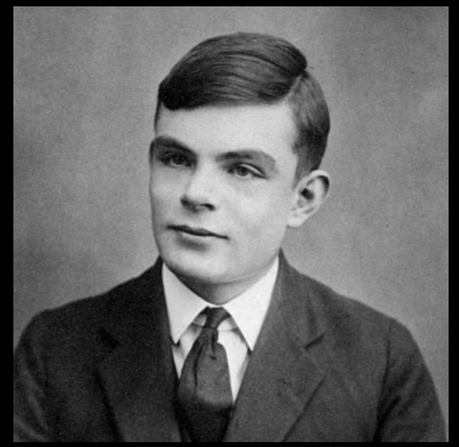
- "A symbolic analysis of replay and switching circuits" (Shannon, 1938)
  - Master's thesis of the century
- Electronic machine could express in terms of Boolean equations with closed and open states of a circuit.
  - The programming as *a problem of formal logic* rather than of arithmetic.
- *"Shannon had injected a subject of purely academic interest into the world of practical machinery."*



Claude Elwood Shannon
(1916–2001)

The father of artificial intelligence 🇬🇧

- Alan Turing (1912–1954)
  - *"Any explicitly stated computational task could be performed by a machine in possession of an appropriate finite set of instructions."*
  - Church–Turing thesis* — *"Anything that can be computed algorithmically can be computed by a Turing machine."*
- On the relationship between human thought and machine thought
  - Turing test (Turing, 1950) to distinguish the answers of humans or machines



*https://en.wikipedia.org/wiki/Church–Turing_thesis

Repercussions of the day

- Warren McCulloch and Walter Pitts (1943) were working on *neural networks*.
  - computations with a suitable finite networks of neurons
  - as a Turing machine (McCorduck 1979)
- John von Neumann (1903–1957)
  - *A stored program* housed in the computer's internal memory
  - The breakthroughs for *assemblers and compilers*, pursuing with the analogies between the brain and computing machines.
  - Worked on the Manhattan Project

Von Neumann's wartime
Los Alamos ID badge photo
(1942–1946)

Von Neumann in the 1940s

Programs of the Dartmouth tetrad

- Now, back to Dartmouth in 1956, we will work-through the three agenda of the four:
  - *"Programs for Problems"* by Allen Newell and Herbert Simon
  - Marvin Minsky and his students
  - *"Lists and Logics"* by John McCarthy

Programs for problems

- Thinking about thinking machines for which they proposed their first program, *Logic Theorist* by Newell and Simon.
- Logic Theorist could actually prove Theorem 2.01 taken from Whitehead and Russell's *Principia* on the JOHNNIAC computer.
- However, *the Journal of Symbolic Logic* declined to publish an article *co-authored* by the Logic Theorist in which this proof was reported (McCorduck, 1979).
- Newell and Simon showed that if they removed the record of previous theorems the Logic Theorist could not solve problems, while humans may can do. This might be the first attempt to see if machines do what humans do.

Marvin Minsky and his students

- Solving analogies of a visual sort (Evans, 1968)
  - The program ANALOGY is shown "A is to B", then to pick the picture out of five which indicates the relations that obtain between "C" and "D."
- STUDENT to solve the problems in high school mathematics book (Bobrow, 1968)
  - *"The gas consumption of my car is 15 miles per gallon. The distance between Boston and New York is 250 miles. What is the number of gallons of gas used on a trip between New York and Boston?"* (Bernstein, 1981, p.113)

Daniel G. Bobrow, Ph.D.
(1935–2017)

Lists and logics

- *LISP*, a programming language that the most widely used in the field of AI in 1960s (McCarthy et al., 1962; Foster, 1967).
  - Both programs and data are structured as lists, or a list of lists.
- The *"intelligence are reduced to a set of logical relationships or axioms that can be expressed precisely in mathematical terms."*
  - *The formalization of concepts: causality, ability, and knowledge* (McCarthy and Hayes, 1969)
- McCarthy has adhered to the general program to model *common-sense reasoning*.

ref. Visual Commonsense Reasoning (Zellers et al., 2019)

Other milestones in 60-70s

- Conversational models
- Kenneth Colby (Standford) and Joseph Weizenbaum (MIT) wrote programs that modeled two-agent dialogues.
- In the light of Turing's famous imitation game, it was a tempting goal to pursue.

Colby's personality chatbot

- Colby's chatbot have psychiatric disorders and so the agent is given by a paranoid personality with an obsession with Mafia:

Why are you in the hospital?

I SHOULDN'T BE HERE.

Who brought you here?

THE POLICE.

What trouble did you have with the police?

COPS DON'T DO THEIR JOB.

Boden, 1977; Colby, 1975

Weizenbaum's ELIZA, the therapist

- Weizenbaum's chatbot was imitating a *Rogerian therapist*, who is performing non-directive but supportive therapy sessions.

Men are all alike.

IN WHAT WAY?

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE?

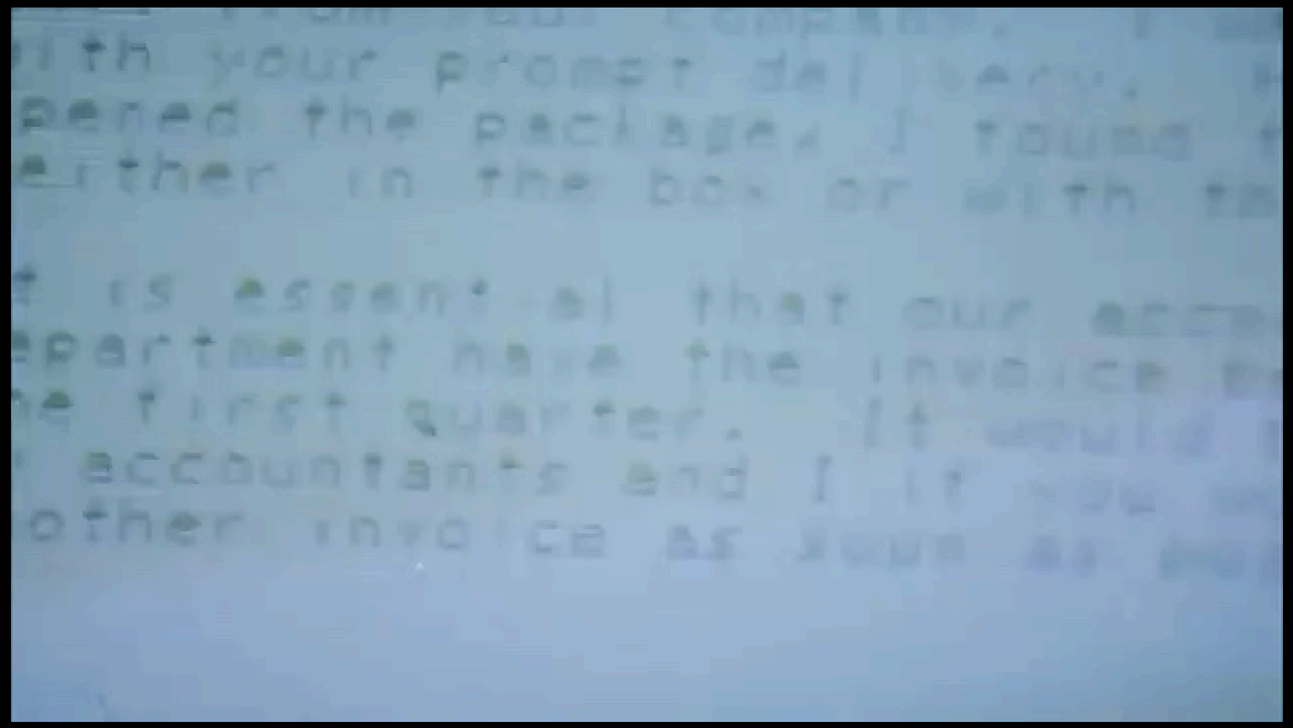
Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE.

Boden, 1977; Weizenbaum, 1966

Weizenbaum's ELIZA, the therapist

- Before ChatGPT, there was ELIZA: the computer therapist chatbot written by Joseph Weizenbaum at MIT's AI Lab in 1966.



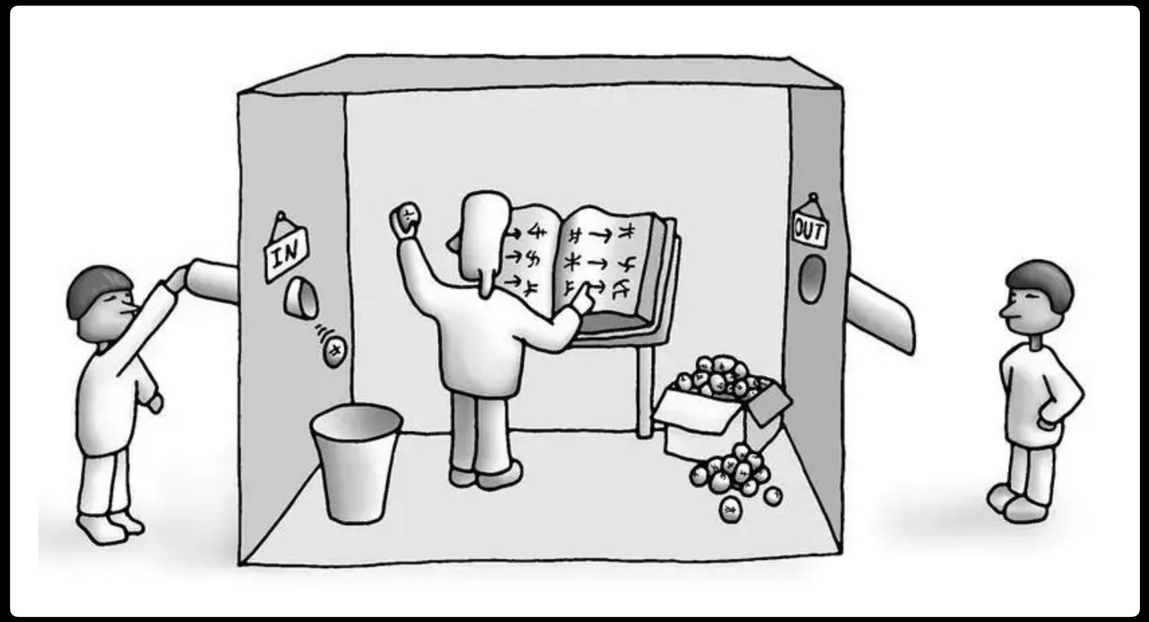
Boden, 1977; Weizenbaum, 1966; Emma Goldman (2017)

Unending debates

- ELIZA depended on *pattern-matching* techniques and sometimes responds with content-free rules, which may be aligned with the *Rogerian* therapist.
- Colby and Weizenbaum initially collaborated, but it diverged in the attitudes toward AI. Weizenbaum posed rather to be practical, aside from the field of AI.
- Colby was a *true believer* who thought that AIs would play an important role in the treatment research of mental illness.
- The debates on the roles and values of AI are still incurring.

Searle's Chinese room

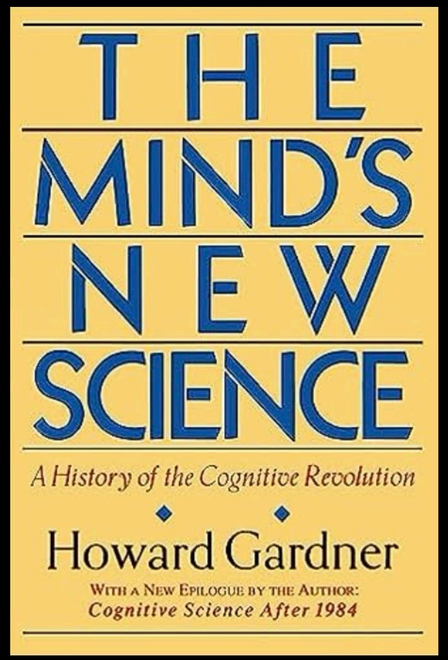
- A *thought experiment* to counter to the Turing test and against the *functionalism* and *computationalism*.
- How do we define "understand" and "intentionality" of agents?

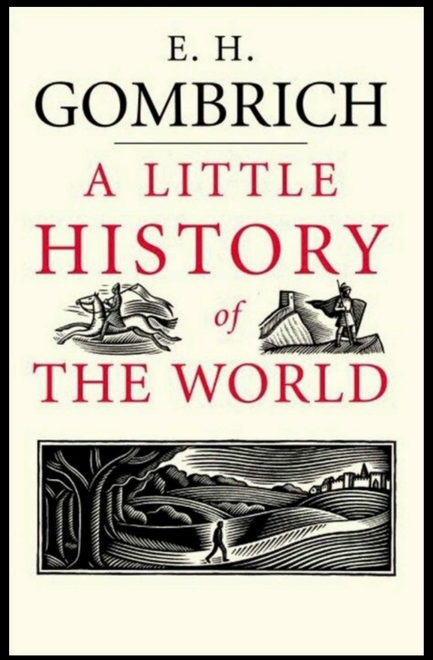


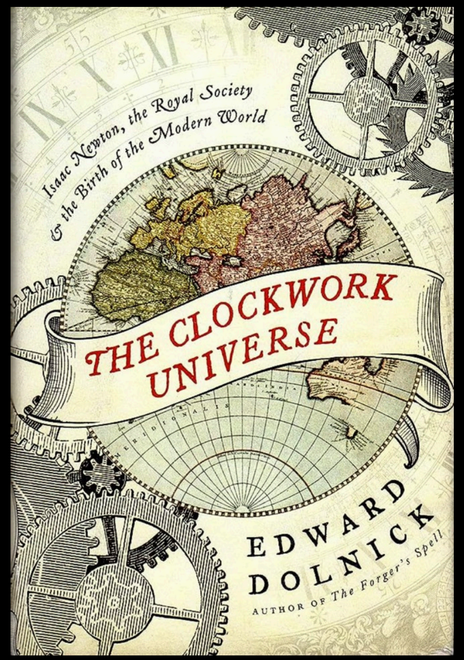
"Minds, Brains, and Programs", John Searle (1980); source: wikicommons

Inspiring books

If your curiosity isn't quite satisfied yet…





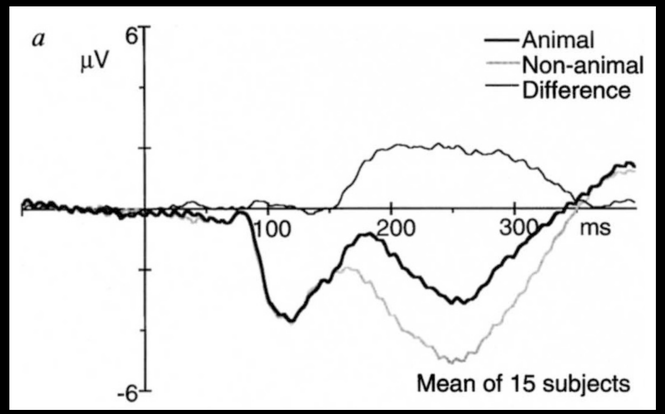


Computer vision

This section is inspired by Fei-Fei's invited talk at NeurIPS 2024.

Speed of objects categorization

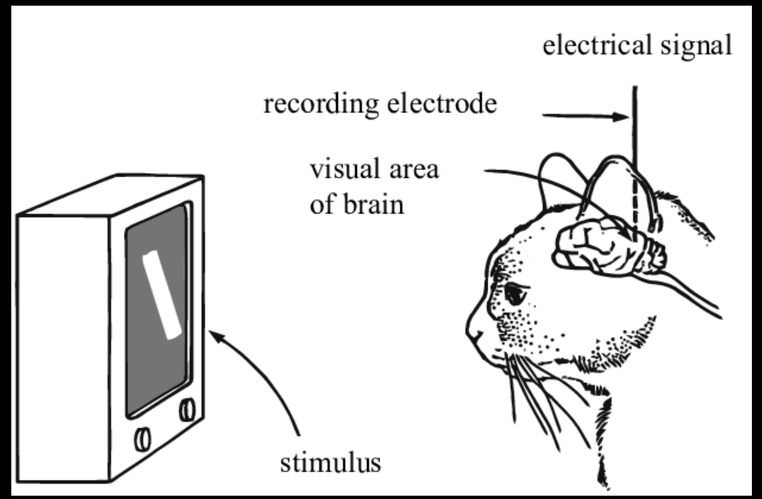
- Object categorization occurs as early as 150 ms (Thorpe et al., Nature 1996).
  - Observing the event-related potentials (ERPs) with a go/no-go categorization task.
- The complexity of the visual recognition pipeline may be summarized as follows:
  - Retinal encoding (0-20 ms)
  - Optic nerve transmission (20-30 ms)
  - Early visual cortex processing (30-70 ms)
  - High-level visual processing (70-120 ms)
  - Categorization response (120-150 ms)



Thorpe et al., 1996

Individual cells respond to certain shapes

- In 1959, Hubel and Wiesel discovered specialized cells in the visual cortex.
- Some cells respond only to horizontal lines, others only to vertical lines.
- Their groundbreaking research earned them the *Nobel Prize in Physiology or Medicine in 1981*.

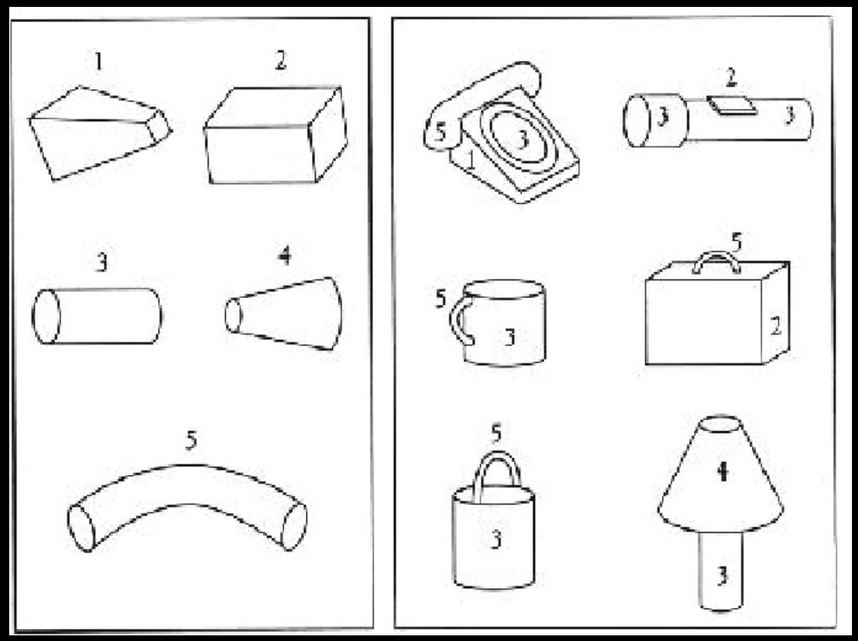


Credit: Nguyen et al., 2019

Hubel & Wiesel, "Receptive fields of single neurones in the cat's striate cortex." J Physiol. 1959.

Early models for object recognition

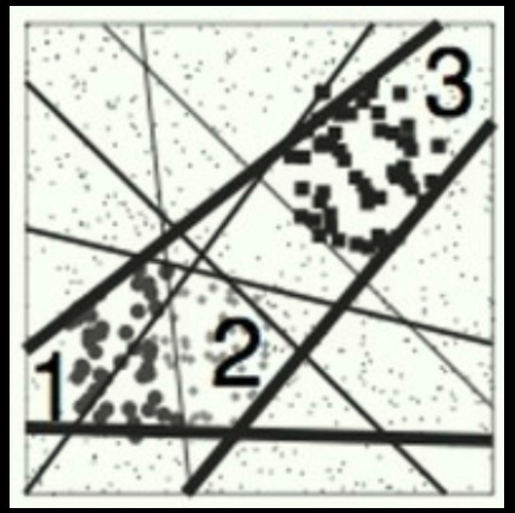
- In psychology, *Geons* refer to simple 2D or 3D forms.
- Recognition-by-components (RBC) theory posits that the geons serve as the primary components of objects (Biederman, 1987).
- ACRONYM (Brooks et al., IJCAI 1979)
- Representing shape for recognition (Marr & Nishihara, 1978)

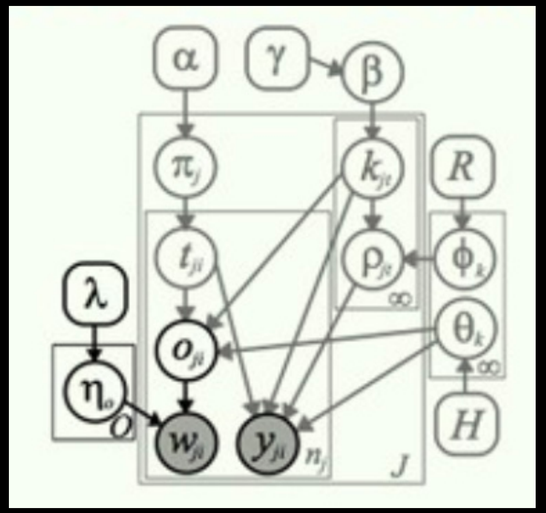


Credit: Biederman, 1990

Early models for object recognition (Cont'd)

- Bag-of-words (Felzenszwalb et al., 2000; Fergus et al., 2003; Fei-Fei et al., 2003)
- Boosting (Chen et al., 2006; Zhu et al., 2007)
- Non-parameteric Bayes (Viola & Jones, 2001; Torralba et al., 2004)

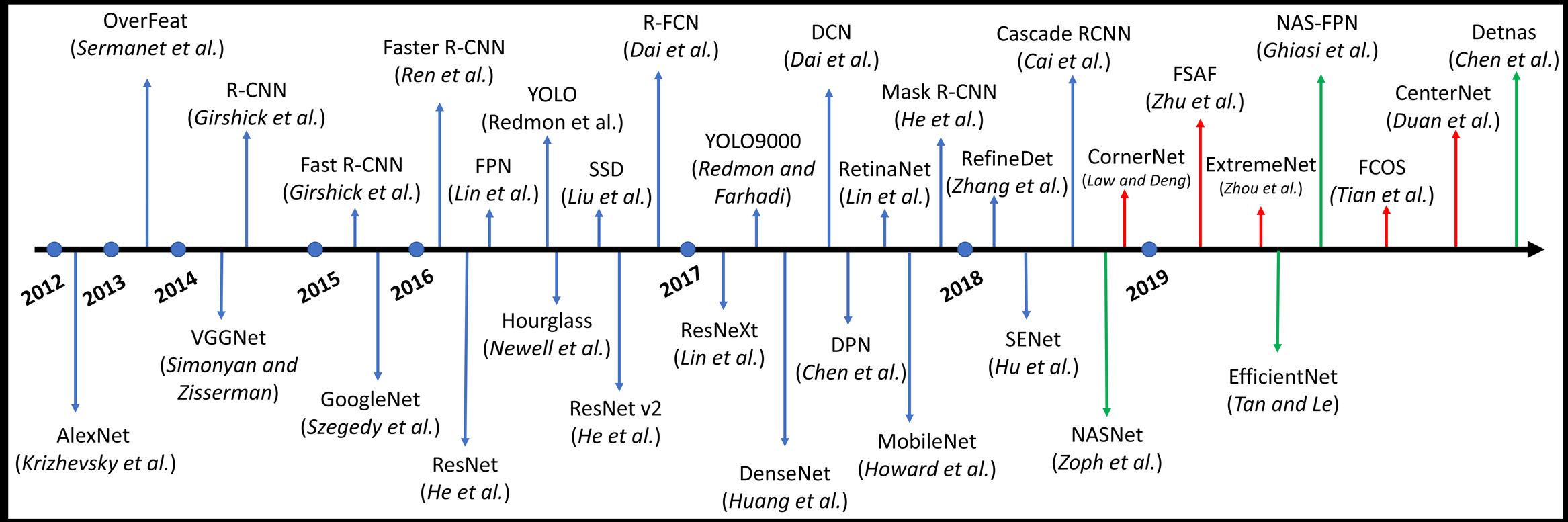




Object classification datasets

- SUN 131 K (Xiao et al., 2010)
- LabelMe 37 K (Russell et al., 2007)
- PASCAL VOC 30 K (Everingham et al., 2006-2012)
- Caltech101 9K (Fei-Fei, Fergus, & Perona, 2003)
- ImageNet *15 M* images w/ 22k categories (Deng et al., CVPR 2009)

Challenges drive progress



Wu et al., "Recent advances in deep learning for object detection", Neurocomputing 2020.

Multimodal deep learning

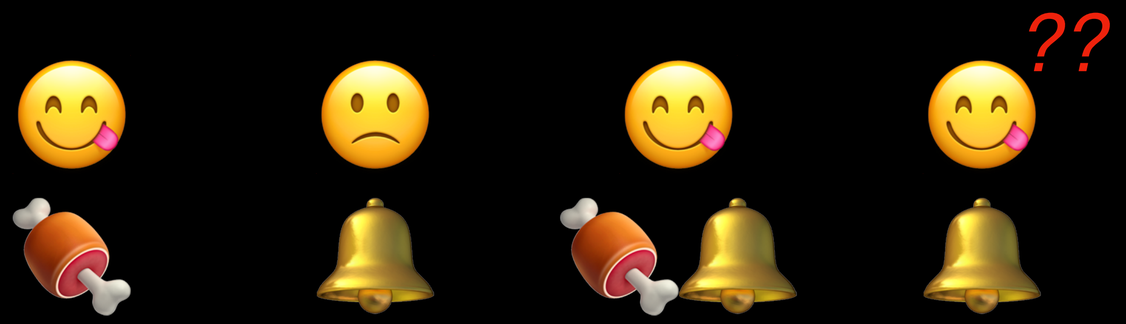
McGurk effect

- McGurk and MacDonald (Nature, 1976) showed a multi-sensory illusion highlighting the quality difference of bimodal stimuli. *"Hearing Lips and Seeing Voices"*
- A vision of /ga/ with a voice of /ba/ *is perceived as* /da/ by most subjects.
- Vision "provides information on the place of articulation and muscle movements, which can help to *disambiguate* between speech with similar acoustics (e.g., the unvoiced consonants /p/ and /k/)." (Summerfield, 1992)
- *Ref. the ventriloquism effect*

https://www.youtube.com/watch?v=2k8fHR9jKVM

Binding problem

- *"The binding problem is from how mammals (particularly higher primates) generate a unified perception of their surroundings from electromagnetic waves, chemical interactions, and pressure fluctuations that forms the physical basis of the world around us. (rephrased)"*
- Classical conditioning, *Pavlov's dog*



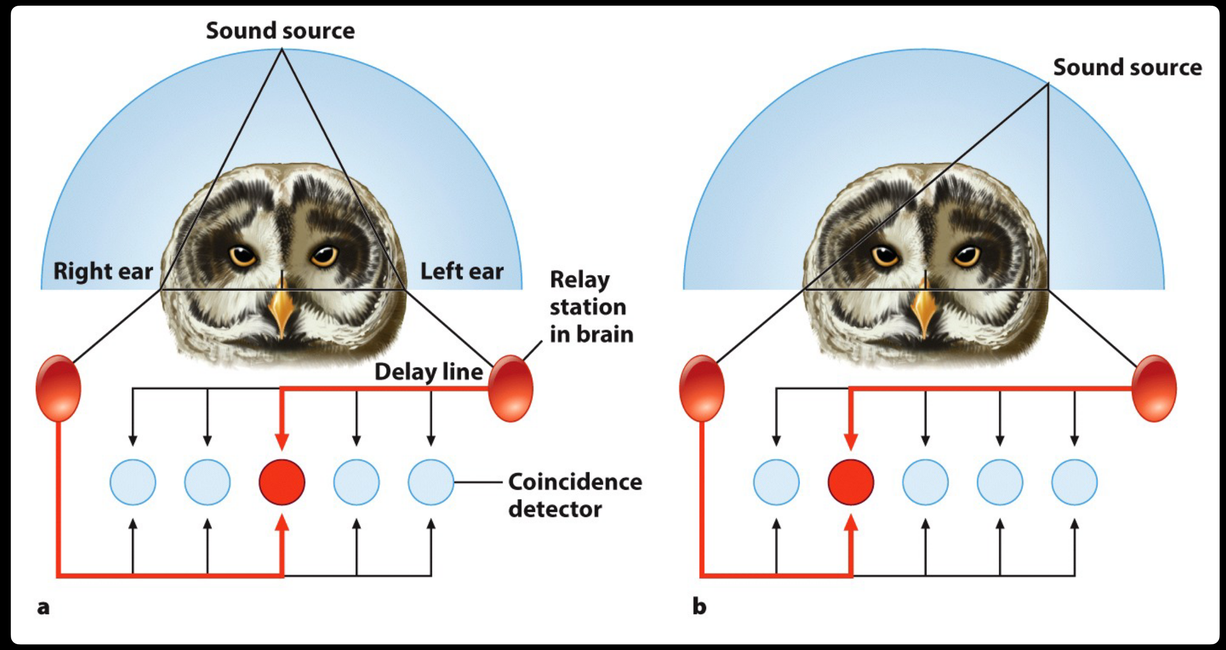
https://en.wikipedia.org/wiki/Multisensory_integration#Binding_problem

Timing is everything

- Timing is critical to the binding problem (Vroomen & Keetels, 2010).
- In physiology, the binding problem is generally referred to temporal synchrony:
  - the coincidence of neural firing,
  - frequent patterns in large scale neural activities.

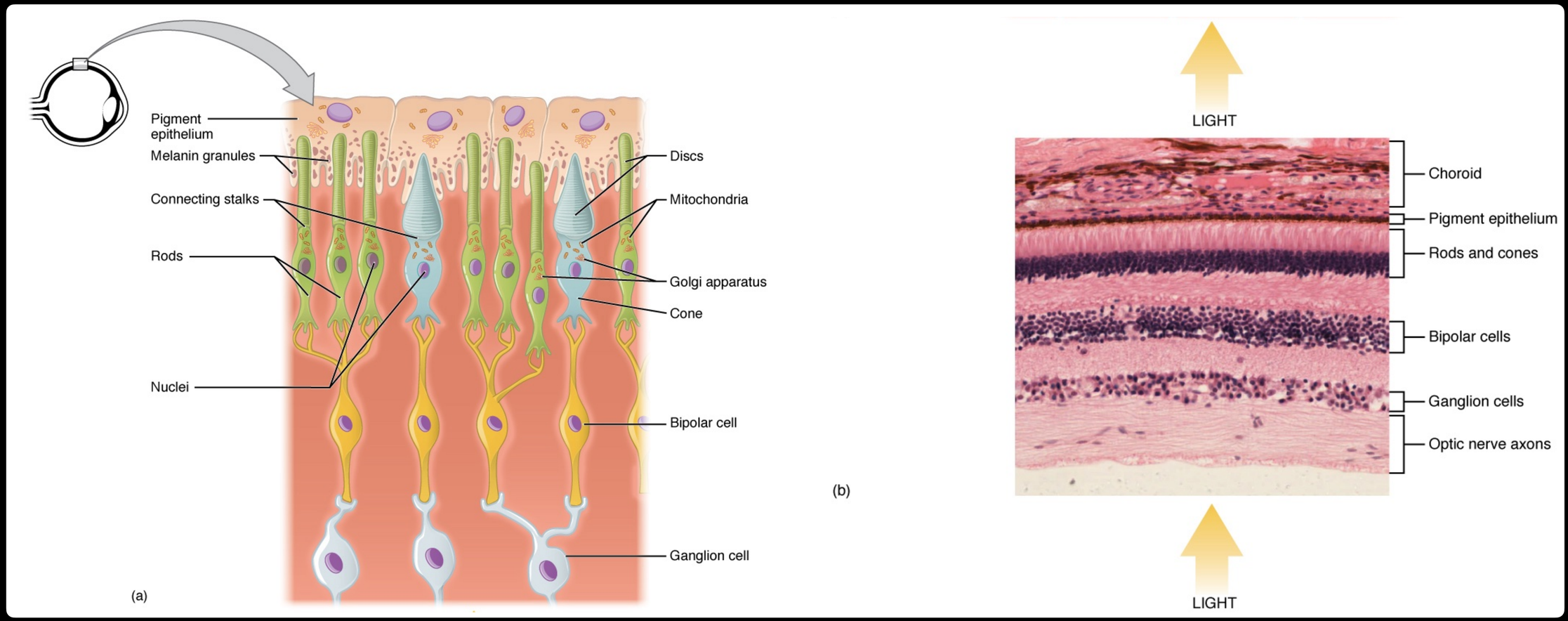
Sound localization in owls

- A example is the *coincidence detector* of barn owls (Carr & Konishi, 1990).
- Using inter-aural time and the difference in sound's intensity at two ears, the asymmetric cues are exploited to localize the source of sound in dark.



Neuroscience, 3rd Ed. (Bear et al., 2006)

Photoreceptor cell



OpenStax College - Anatomy & Physiology, Connexions Web site. http://cnx.org/content/col11496/1.6/

Representation binding

- In deep learning, learning joint representation to solve the binding problem in multimodal learning tasks, *e.g.*, visual question answering (will be discussed).
- *Mid-level* embedding for each modality is crucial (Ngiam et al., 2011).
- Large models for each modality is for state-of-the-art performance (Chen et al., 2022).
- Early multimodal learning methods may include:
  - Bilinear models, cross-attention models, and
  - Self-supervised learning maximizing mutual information.

ChatGPT

A brief of ChatGPT

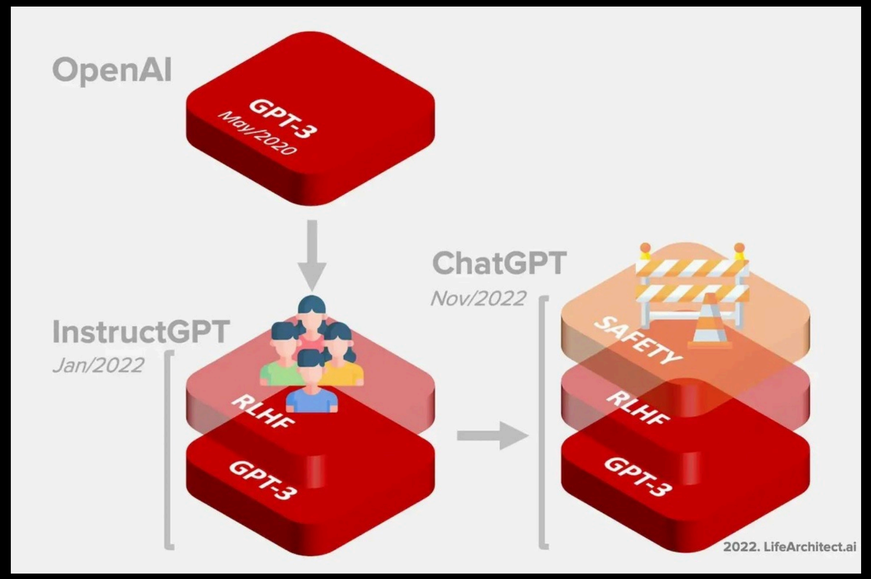
- ChatGPT is a chat system based on Open AI' language model GPT-3, or Generative Pre-training Transformer 3.
- It may respond to inquiries like tourism recommendations, translations, summarizing, and coding. (polishing paragraphs and correcting grammar errors in writing papers!)
- It is known that the datasets cover only data before 2021*.

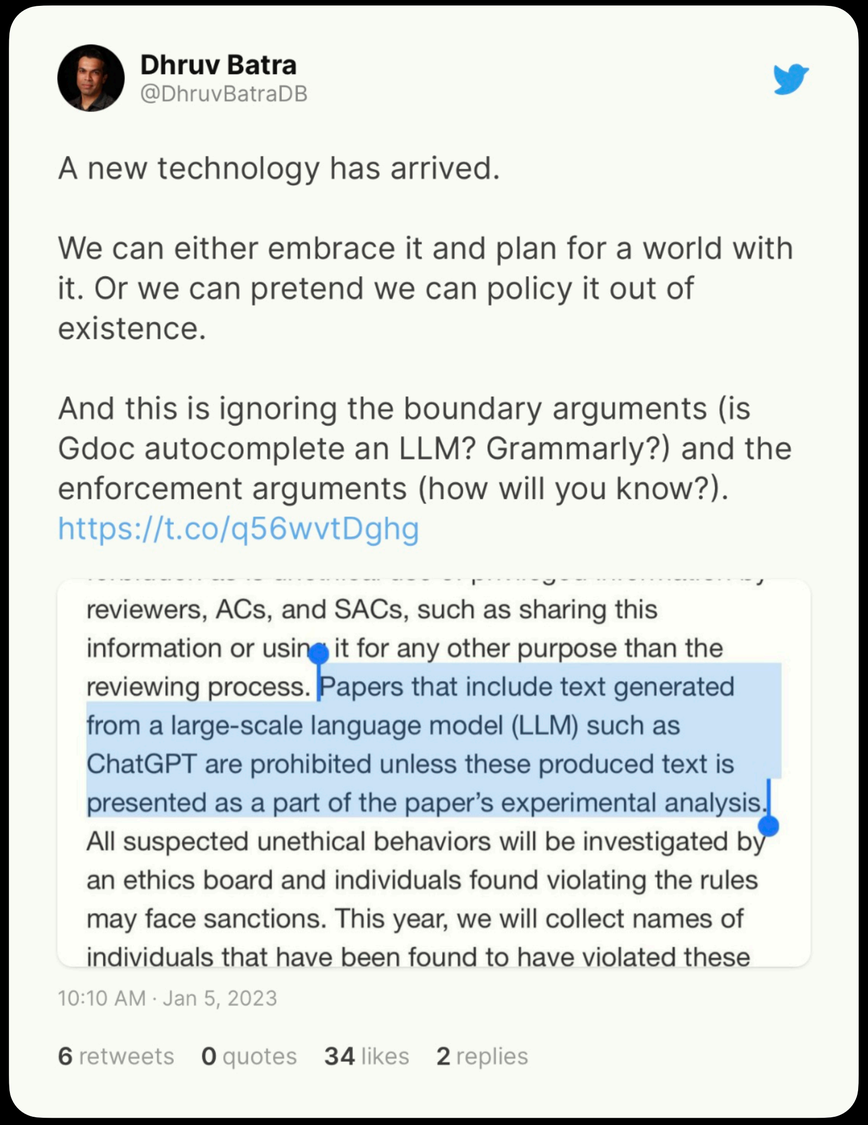
I'm sorry, but I am not able to browse the internet or access current events, as my knowledge is based on the information that was available to me at the time of my training. My knowledge cutoff is September 2021, so I am not able to provide information about events that have occurred after that time.

*ChatGPT as of Jan 6, 2023. GPT-4 Turbo covers up to Apr 2023.

A brief of ChatGPT (cont'd)

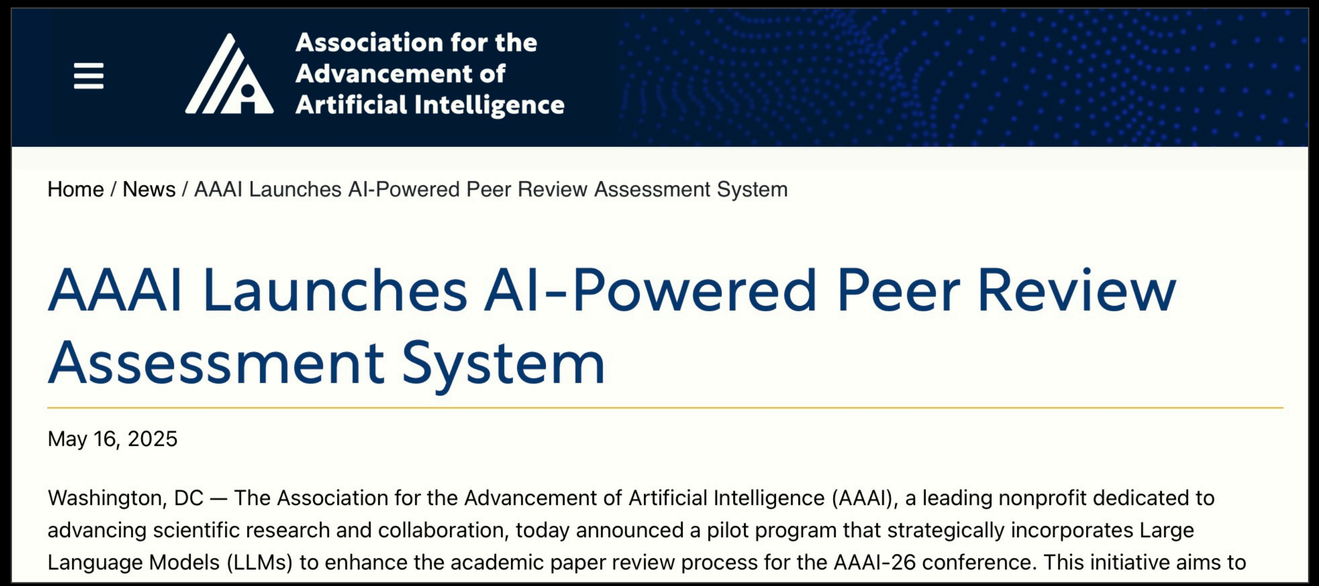
- The concern of plagiarism will be controversial.
  - The CfP of ICML 2023 explicitly mentioned this worry:
- GPT-4o was released in May 2024 and is expected to be multimodal, including sound, video, and image.

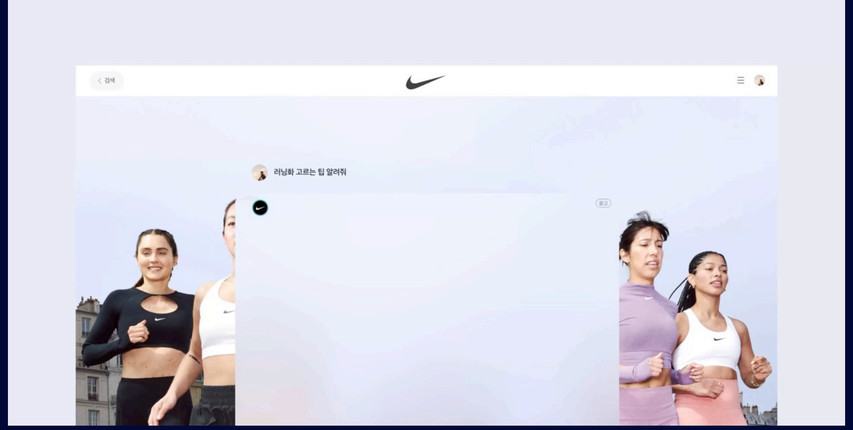




AAAI'26 adopts AI-powered review system

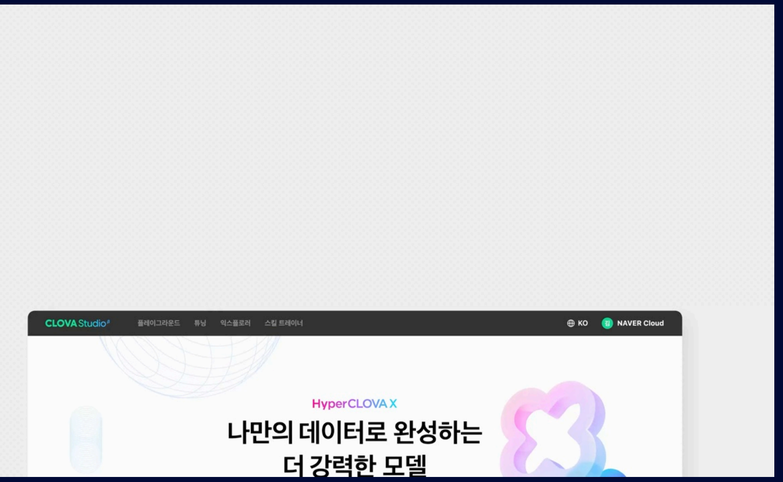
- "Enhancing scientific Review, not replacing human expertise"
  - Supplementary first-stage reviews and discussion summary assistance
- "Preserving human decision-making and scientific integrity"
- "cutting-edge methods with rigorous safeguards"

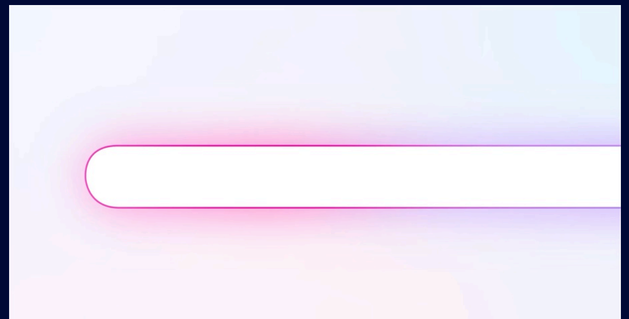




HyperCLOVA X

NAVER Generative AI Model





NAVER AI LAB

NAVER WORKS

Collaboration

WORKS Core
All-in-one collaboration tool
Mail, Message, Drive, Tasks, form

WORKS Drive
Drive for work

coming soon

WORKS On
Video conference for work

AI Product
AI meeting Note
AI Push to talk
AI Builder

Management

Korea Only

WORKS Approval
E-doc approval

WORKS Attendance
Attendance Management System

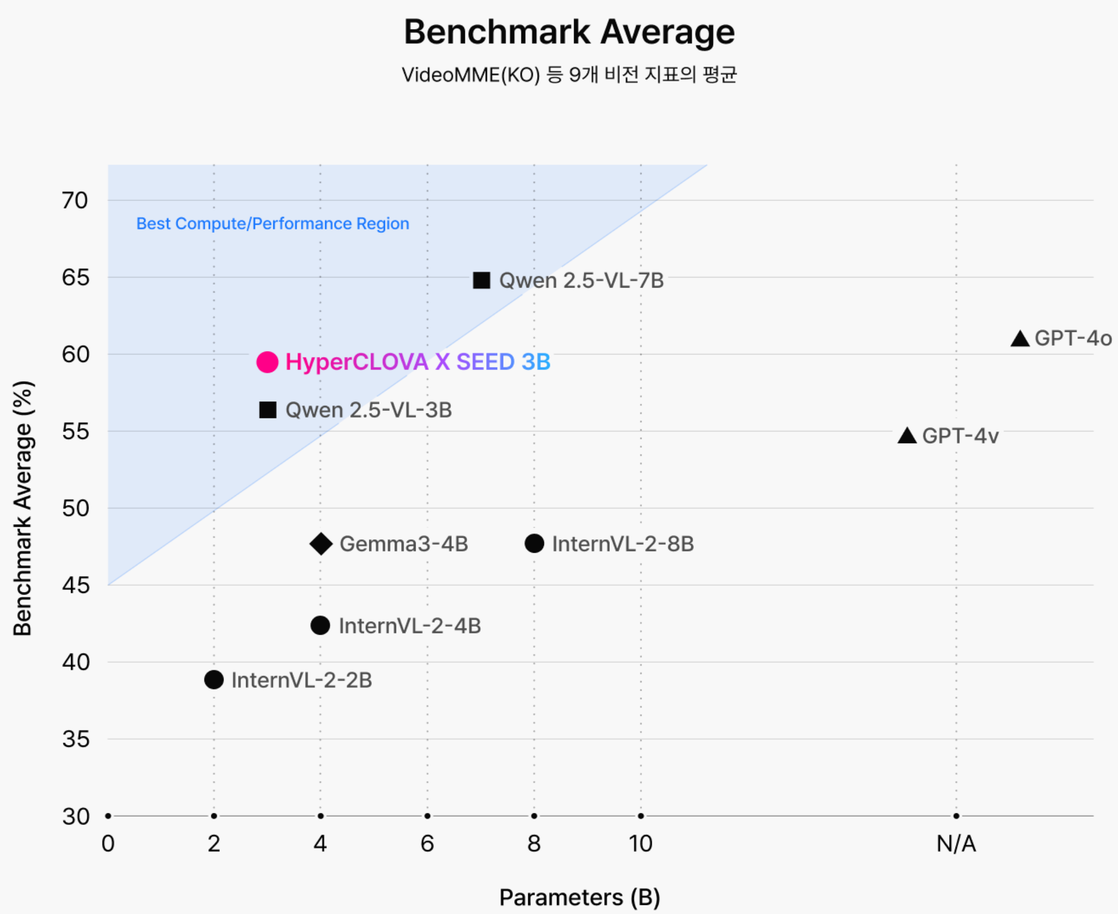
WORKS Payroll
Salary payment

WORKS Finance
Reimbursement/ Accounting

WORKS AI

HyperCLOVA X

NAVER Generative AI Model



Released a lightweight HCX model ‘HCX-SEED’ as a free open-source (‘25.04)

Drive the growth of Korea’s AI ecosystem

- Deployment in NAVER services
- Development of enterprise solutions
- Open-source contributions

NAVER AI LAB

HyperCLOVA X Reasoning Model (Early Preview)

HyperCLOVA X Reasoning delivers world-class QA through test-time compute scaling, matching GPT-4o search in Korean and English.

Simple English QA Score

Search?	Model Type	Score
O	HyperCLOVA X Reasoning	90.1
O	OpenAI gpt-4o-search-preview	90.0
X	OpenAI gpt-4o	38.2
X	HyperCLOVA X Non-Reasoning	4.95

Simple Korean QA Score

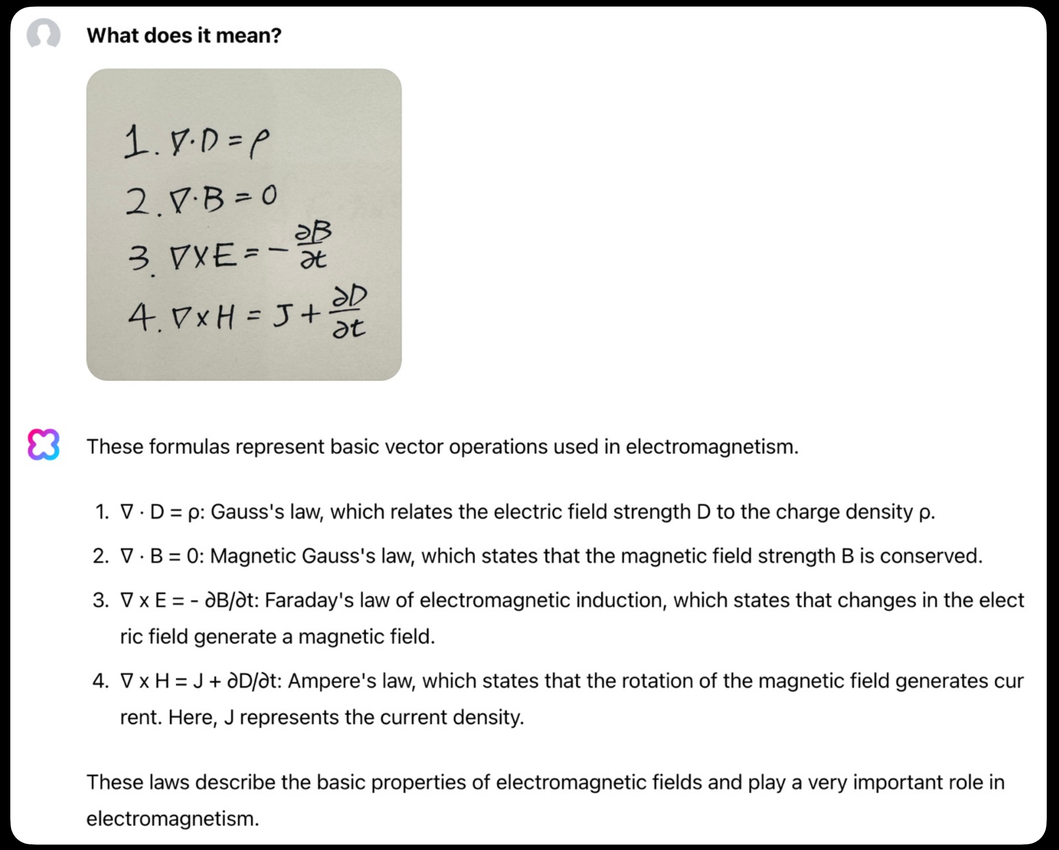
Search?	Model Type	Score
O	OpenAI gpt-4o-search-preview	87.4
O	HyperCLOVA X Reasoning	87.2
X	OpenAI gpt-4o	75.2
X	HyperCLOVA X Non-Reasoning	66.6

HyperCLOVA X Multimodal

HyperCLOVA X Vision fuses Korean language modeling with image understanding, matching or surpassing GPT-4V on key benchmarks to deliver sovereign multimodal intelligence.

K-GED Performance

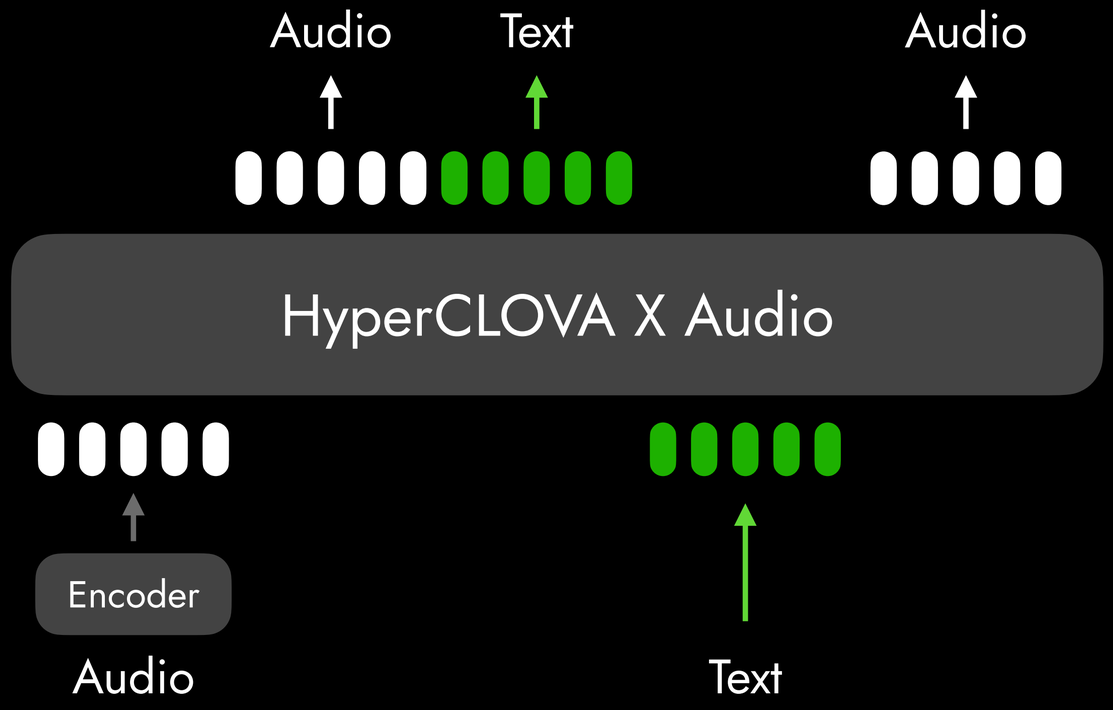
Model	Model Type
HCX-VLM	1240/1480 (83.8%)
GPT-4o	1152/1480 (77.8%)



K-GED Performance is a recall-style graph similarity metric.

HyperCLOVA X Multimodal

HyperCLOVA X Audio recognizes the emotion in your voice and generates expressive, human-level speech.



https://clova.ai/tech-blog/ko-hyperclova-기반-음성-합성-기술-audiollm

Multimodal Generation

Meta Open Arts in Times Square

- *"Meta Open Arts builds community through creativity, cultivating creative engagements that offer new ways of thinking about ourselves, our communities, and the world at large."*[1]

Credit: Tatyana Tenenbaum, courtesy of Times Square Arts. May 2022

[1] https://www.facebook.com/MetaOpenArts/

Endangered species

- *Endangered species*, the residence project for FAIR Artists, are displayed over Times Square by Sofia Crespo from Entangled Others Studio.
- A generative model trained on 10K open-source images of animal and plant species.
- *"The resulting models were then used to generate visual representations of lesser-known species that are critically endangered."*

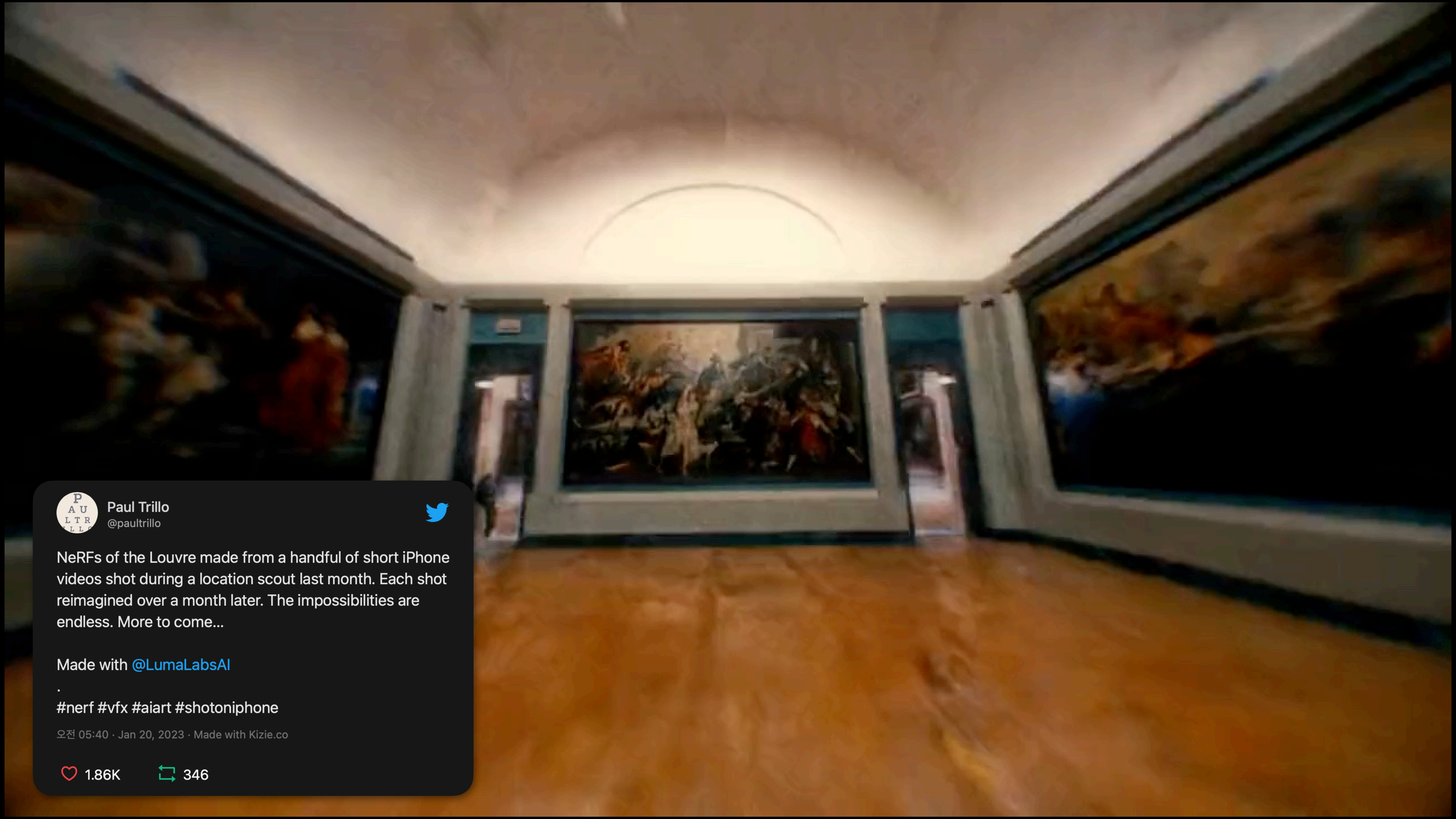
https://entangledothers.studio

"An A.I.-Generated Picture Won an Art Prize. Artists Aren't Happy."

By Kevin Roose, The New York Times (Sept. 2, 2022)

Journey of a lifetime

Luma AI's keyframes feature on Dream Machine 1.5 by @kodykurth



Paul Trillo
@paultrillo

NeRFs of the Louvre made from a handful of short iPhone videos shot during a location scout last month. Each shot reimagined over a month later. The impossibilities are endless. More to come...

Made with @LumaLabsAI
.
#nerf #vfx #aiart #shotoniphone

오전 05:40 · Jan 20, 2023 · Made with Kizie.co

1.86K 346

Higgsfield's 3D Rotation

https://higgsfield.ai (August 2025)

Any questions?

Please don't hesitate to ask question for being a part of this discussion!