ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

CENG 463: Introduction to Natural Language Processing
Introduction Session

Asst. Prof. Cagri Toraman
Computer Engineering Department
ctoraman@ceng.metu.edu.tr

30.09.2025

About the Instructor

Dr. Çağrı Toraman

- 2009, 2011, 2017 B.Sc., M.Sc., Ph.D. Bilkent University Computer Science
- 2017 - 2018 Havelsan Big Data Group, Software Engineer
- 2018 - 2019 University of Central Florida, Postdoctoral Research Scientist
- 2019 - 2023 Aselsan Research Center, NLP Team Lead
- 2024 - .. Middle East Technical University, Computer Engineering Department, Applied NLP Group

Course Details

ODTUClass
https://odtuclass2025f.metu.edu.tr/

Tuesday 11:40 - 13:30
Wednesday 11:40 - 12:30
BMB-4

Programming Assignment: 18% (5-5-8)
Quiz: 20% (6-6-8)
Midterm Exam: 25%
Final Exam: 30%
Active Participation: 7%

Artificial Intelligence

A branch of Computer Science.

Examines how we can achieve intelligent behaviour through computation.

Natural Language Processing

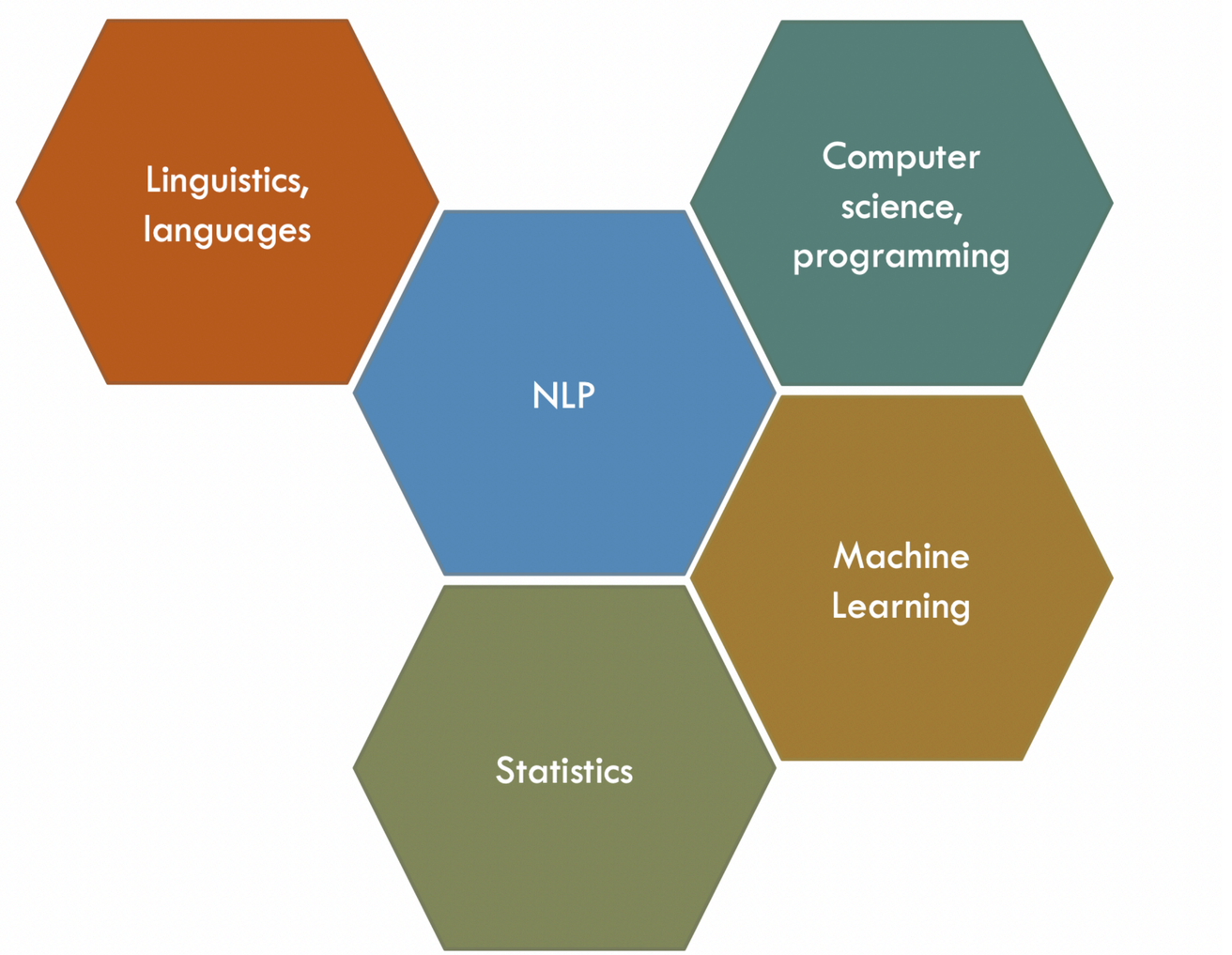
A branch of AI.

Technology to handle human language (usually text) using computers

Human-machine communication: Question answering

Human-human communication: Machine translation

Analyze/understand language: Syntactic analysis


Linguistics, languages
Computer science, programming
NLP
Machine Learning
Statistics

Our Objectives in This Class

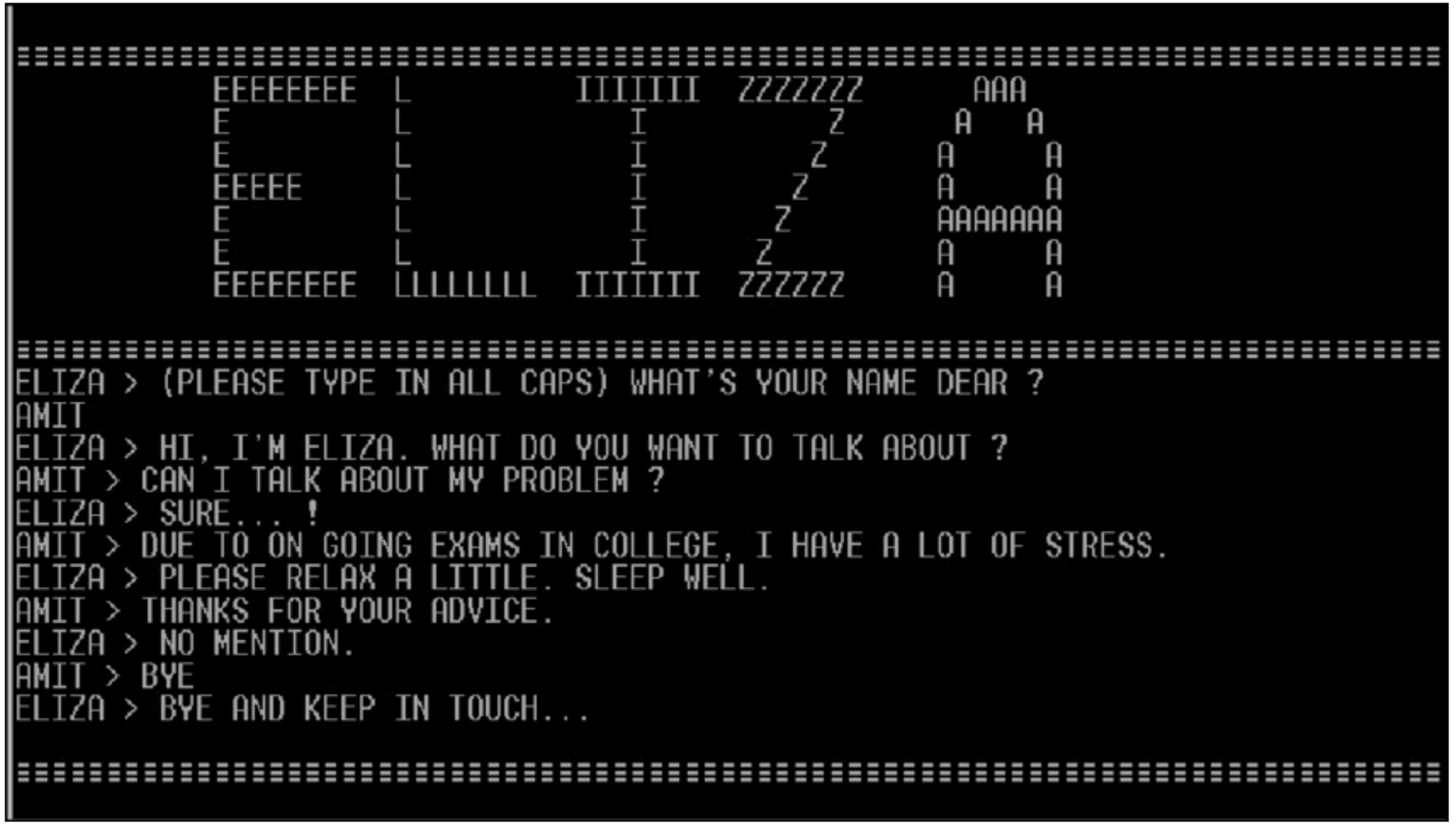
Understand the basic principles and the history of NLP.

Learn the essential elements of NLP including syntax and semantics.

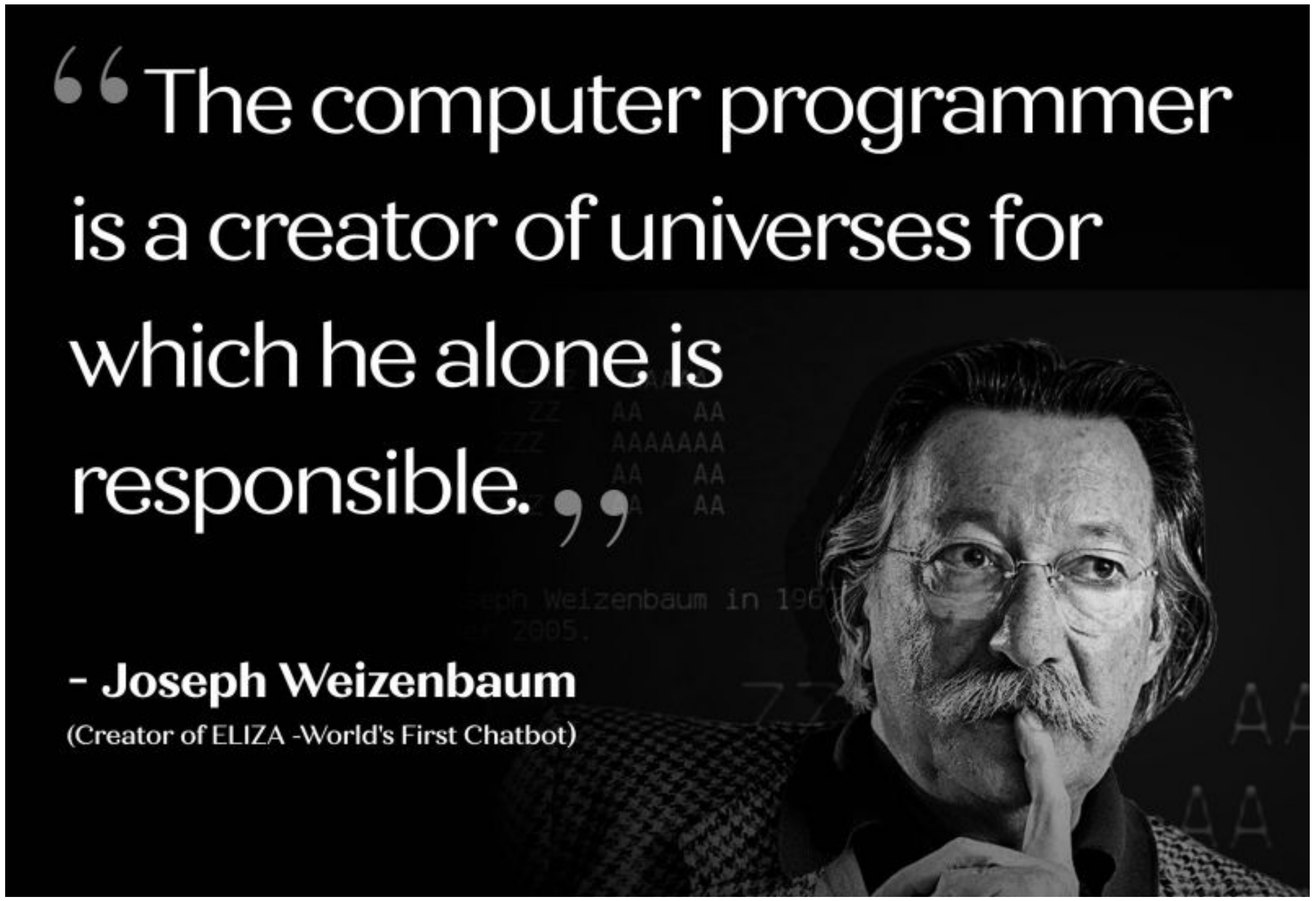
Explore a variety of modern NLP tasks such as text classification, token classification, and training large language models.

Our (brief) journey to AI

1966


EEEEEEEE L IIIIIII ZZZZZZZ AAA
E L I Z A A
E L I Z A A
EEEEE L I Z A A
E L I Z AAAAAAA
E L I Z A A
EEEEEEEE LLLLLLLL IIIIIII ZZZZZZ A A
ELIZA > (PLEASE TYPE IN ALL CAPS) WHAT'S YOUR NAME DEAR ?
AMIT
ELIZA > HI, I'M ELIZA. WHAT DO YOU WANT TO TALK ABOUT ?
AMIT > CAN I TALK ABOUT MY PROBLEM ?
ELIZA > SURE... !
AMIT > DUE TO ON GOING EXAMS IN COLLEGE, I HAVE A LOT OF STRESS.
ELIZA > PLEASE RELAX A LITTLE. SLEEP WELL.
AMIT > THANKS FOR YOUR ADVICE.
ELIZA > NO MENTION.
AMIT > BYE
ELIZA > BYE AND KEEP IN TOUCH...

1966


“The computer programmer is a creator of universes for which he alone is responsible.”
- Joseph Weizenbaum
(Creator of ELIZA -World's First Chatbot)

2010s

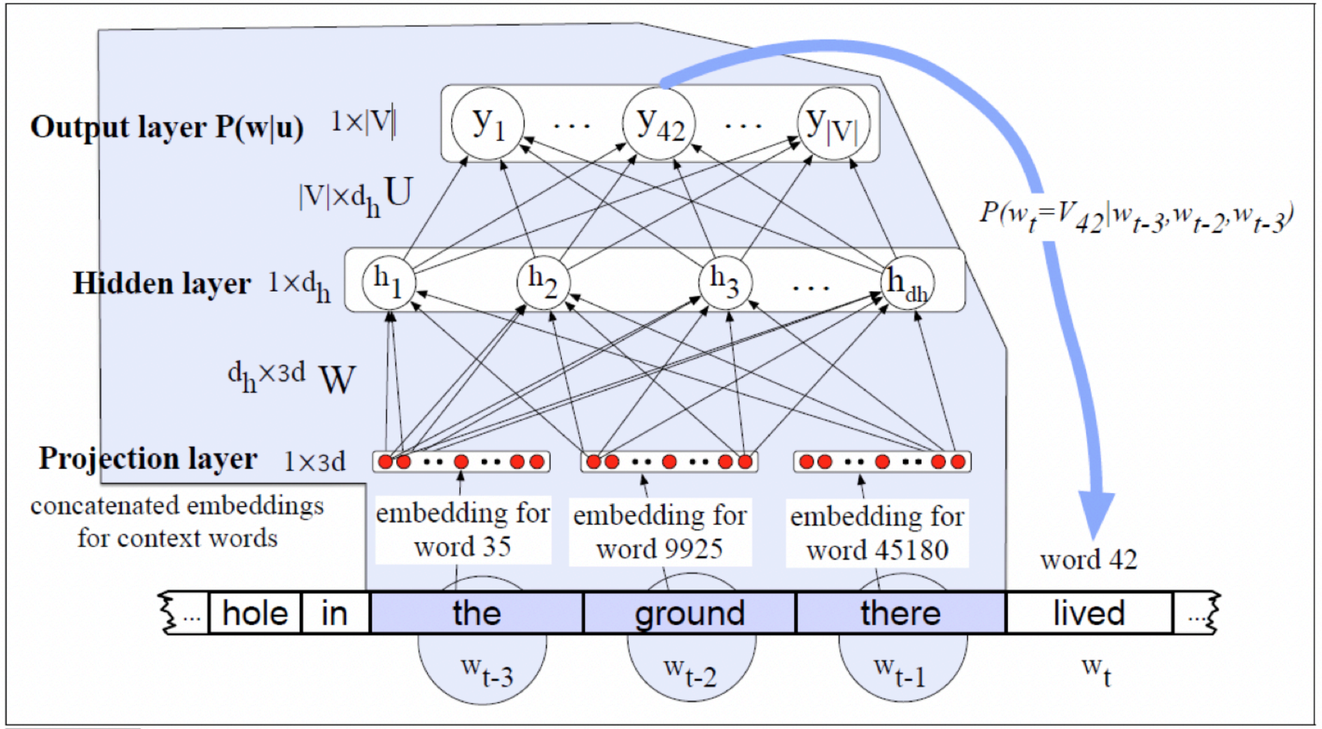
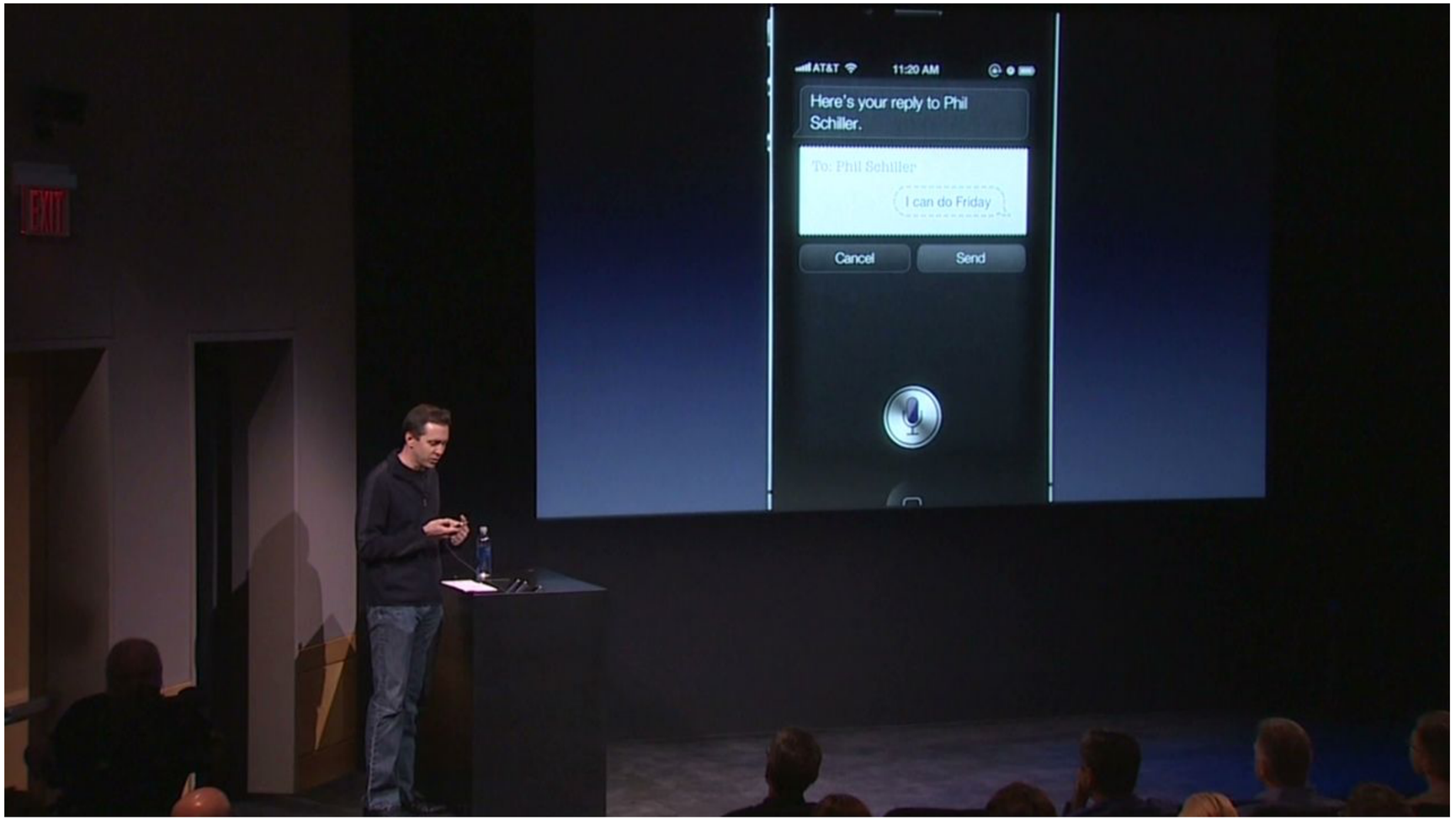


Figure 9.1 A simplified view of a feedforward neural language model moving through a text. At each time step t the network takes the 3 context words, converts each to a d-dimensional embedding, and concatenates the 3 embeddings together to get the $1 \times Nd$ unit input layer x for the network.

2011


AT&T
11:20 AM
Here's your reply to Phil Schiller.
To: Phil Schiller
I can do Friday
Cancel
Send
EXIT

2011

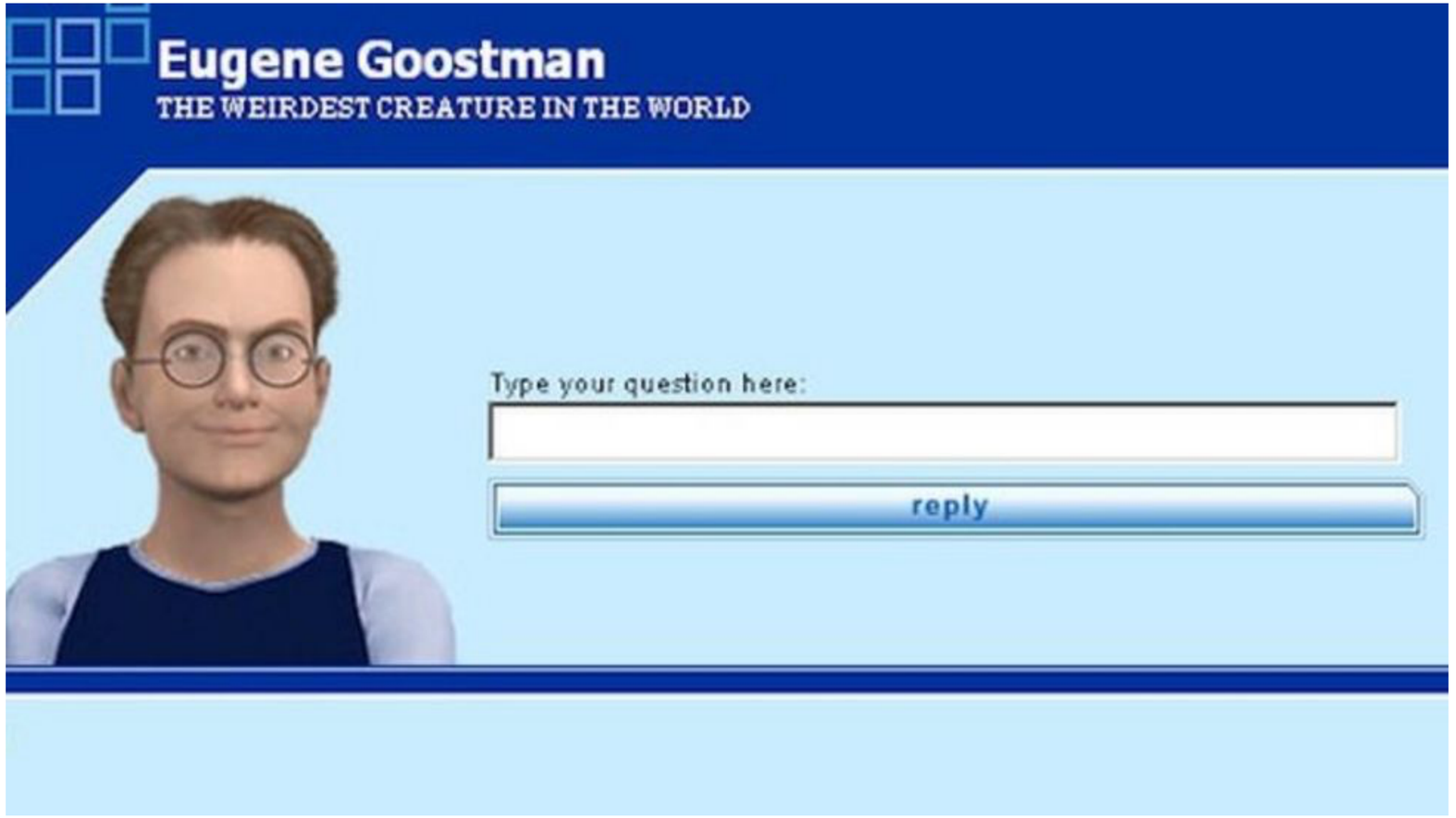
2014

Amazon Echo Is A $199 Connected Speaker Packing An Always-On Siri-Style Assistant

Darrell Etherington / 9:14 AM PST • November 6, 2014

Comment

2014


Eugene Goostman
THE WEIRDEST CREATURE IN THE WORLD
Type your question here:
reply

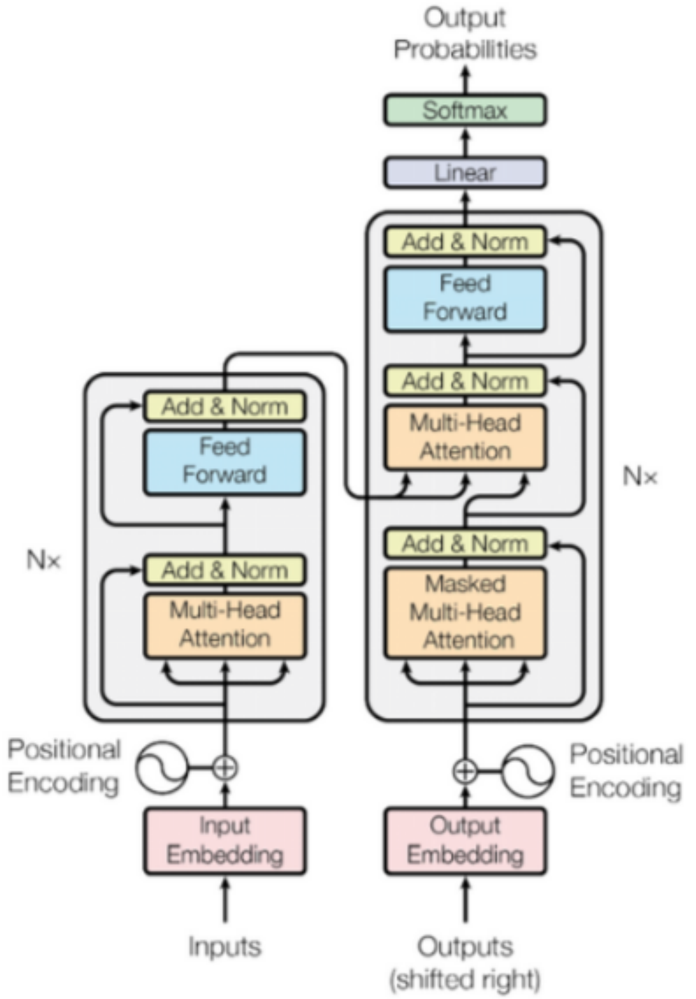
2017


RETHINK YOUR HEROES
A MICHAEL BAY FILM
TRANSFORMERS
THE LAST KNIGHT
#TRANSFORMERS
JUNE

2017

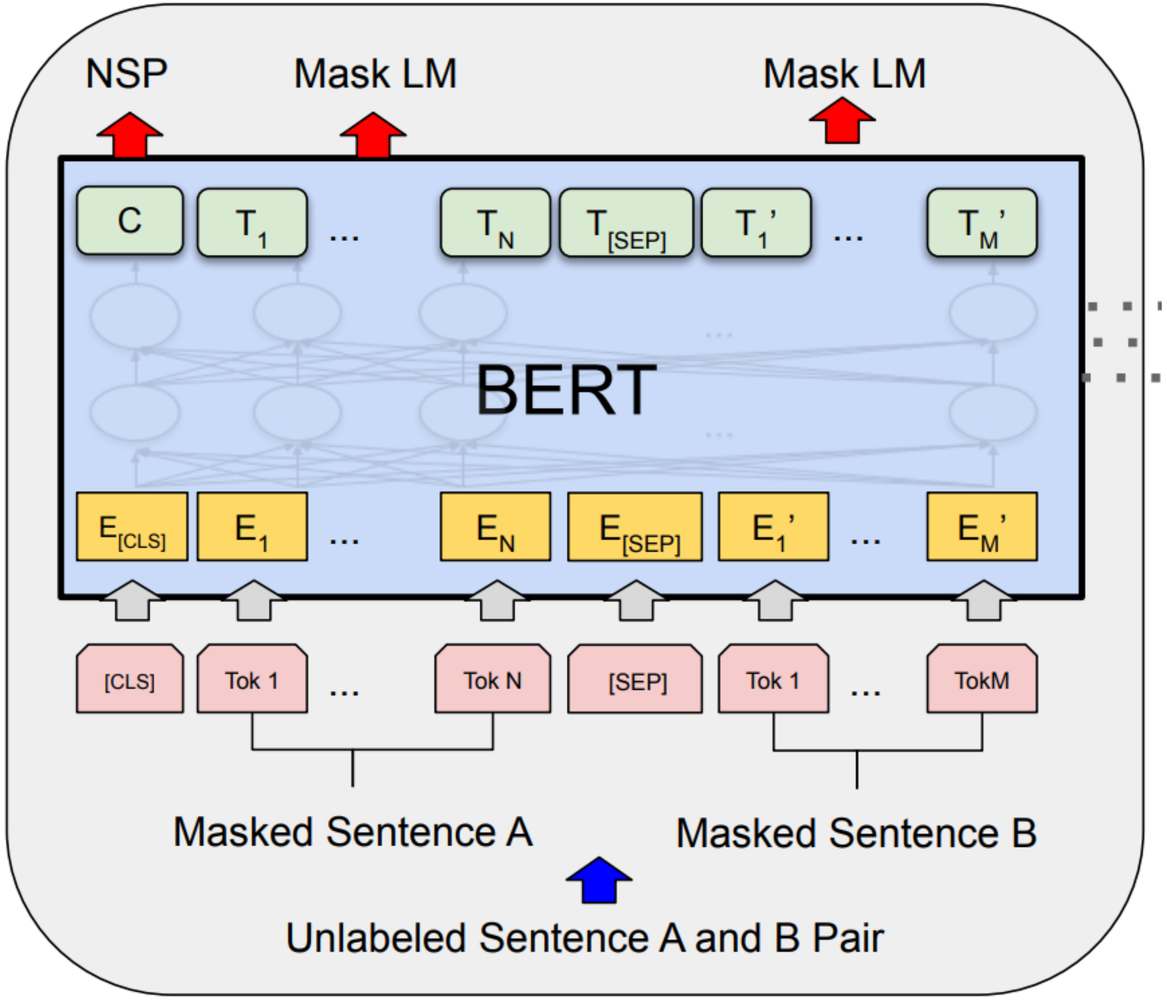
Transformer

Attention Is All You Need



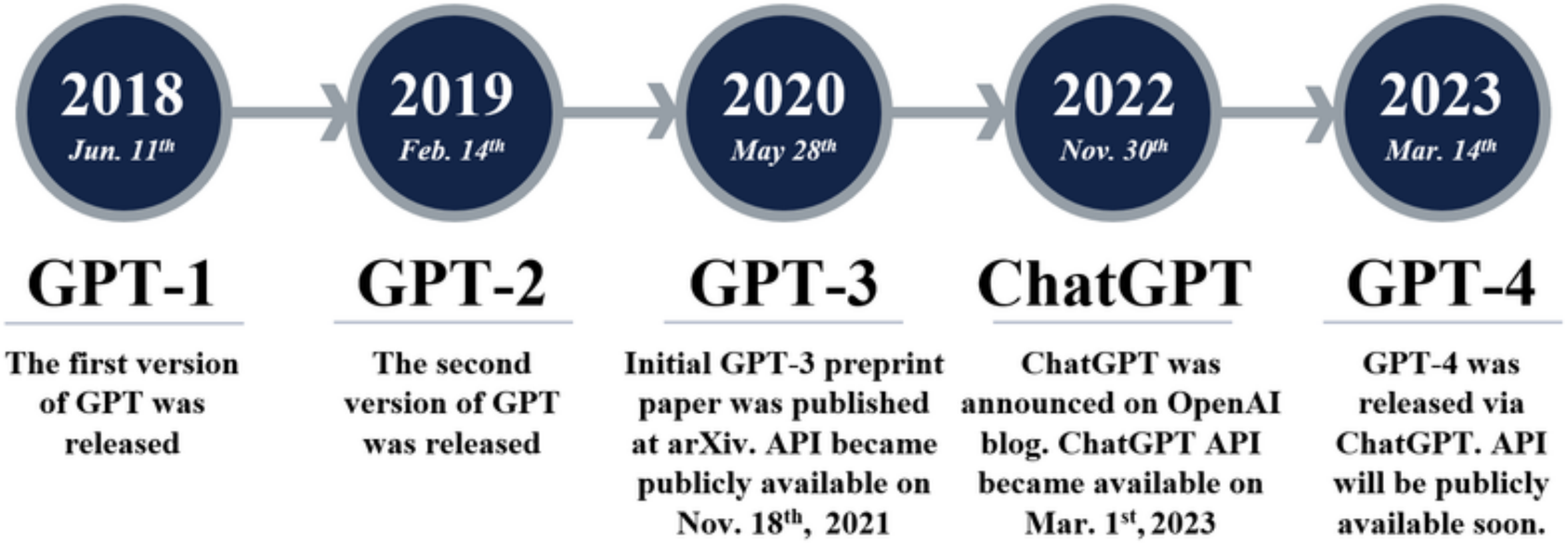
2018

2018



Pre-training

2018 - … Üretken Yapay Zeka (Generative AI)

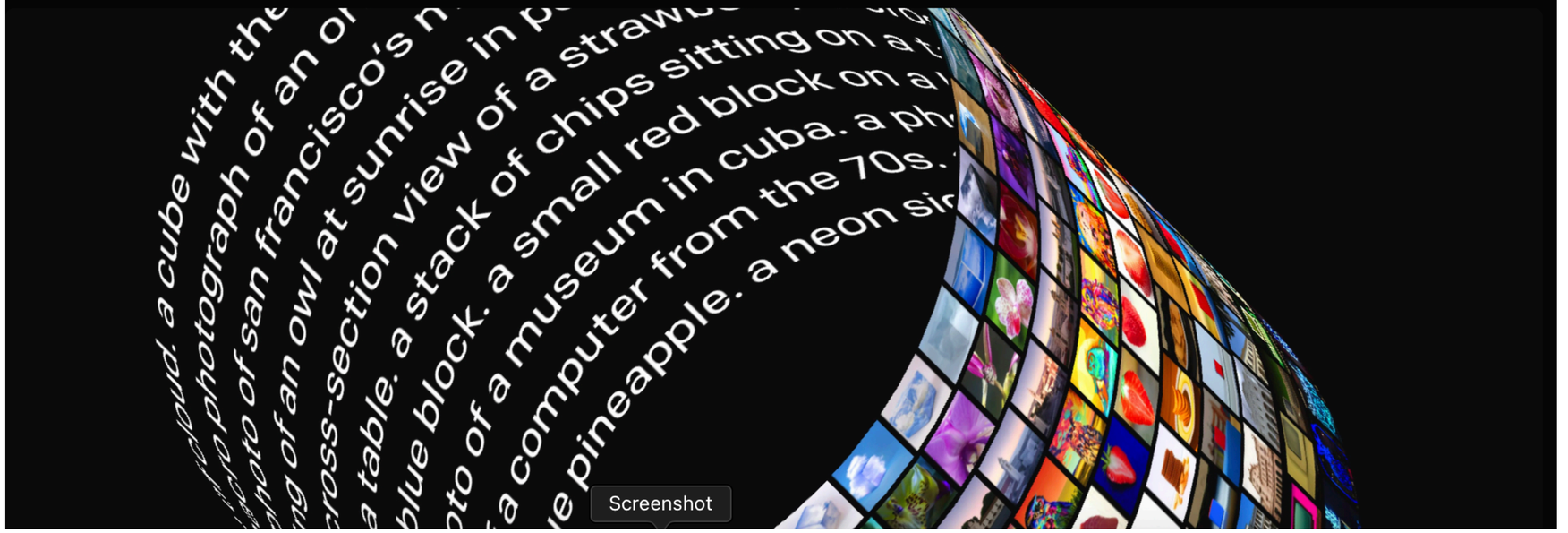


2021

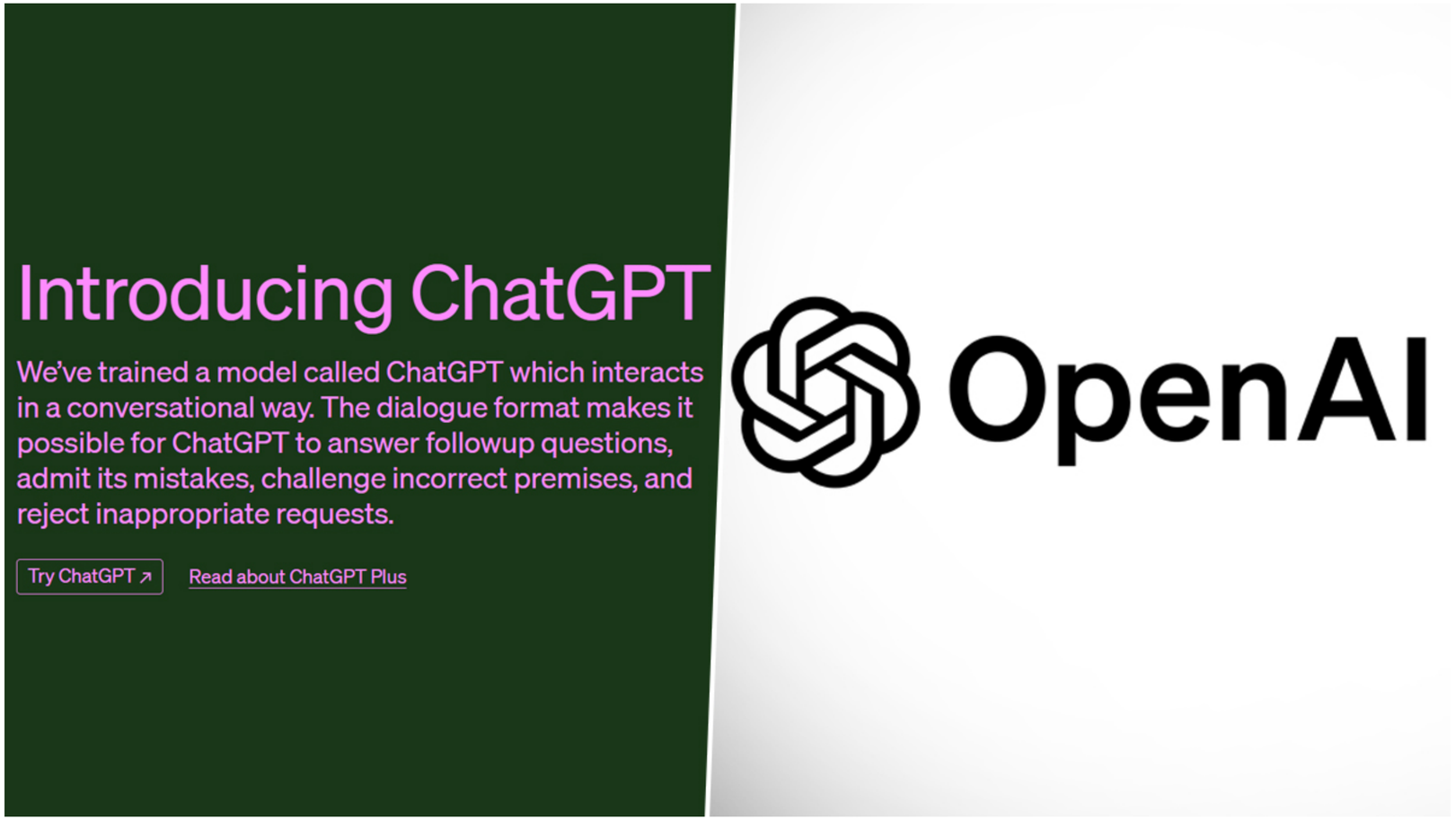
January 5, 2021

DALL·E: Creating images from text

We've trained a neural network called DALL·E that creates images from text captions for a wide range of concepts expressible in natural language.

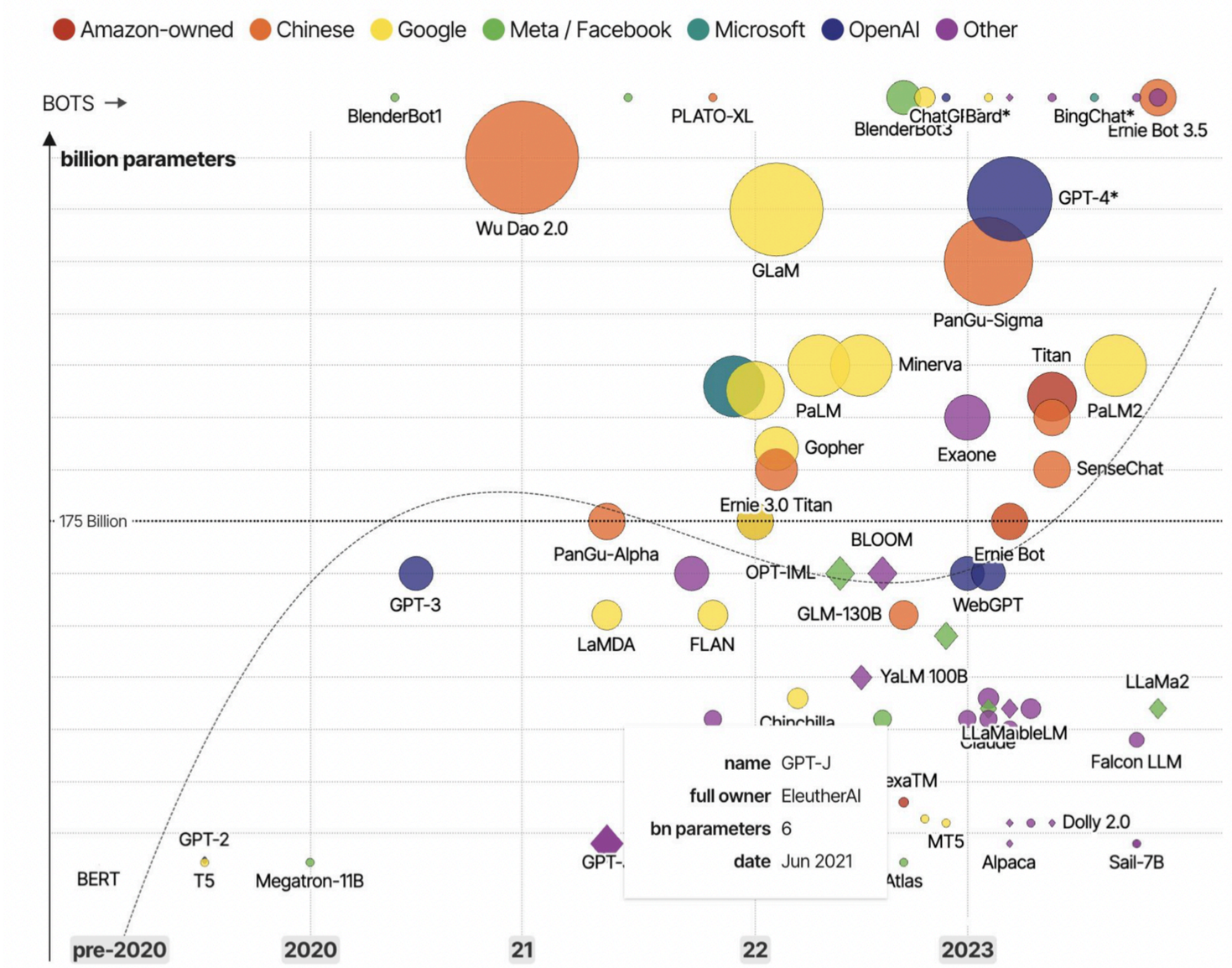


2022

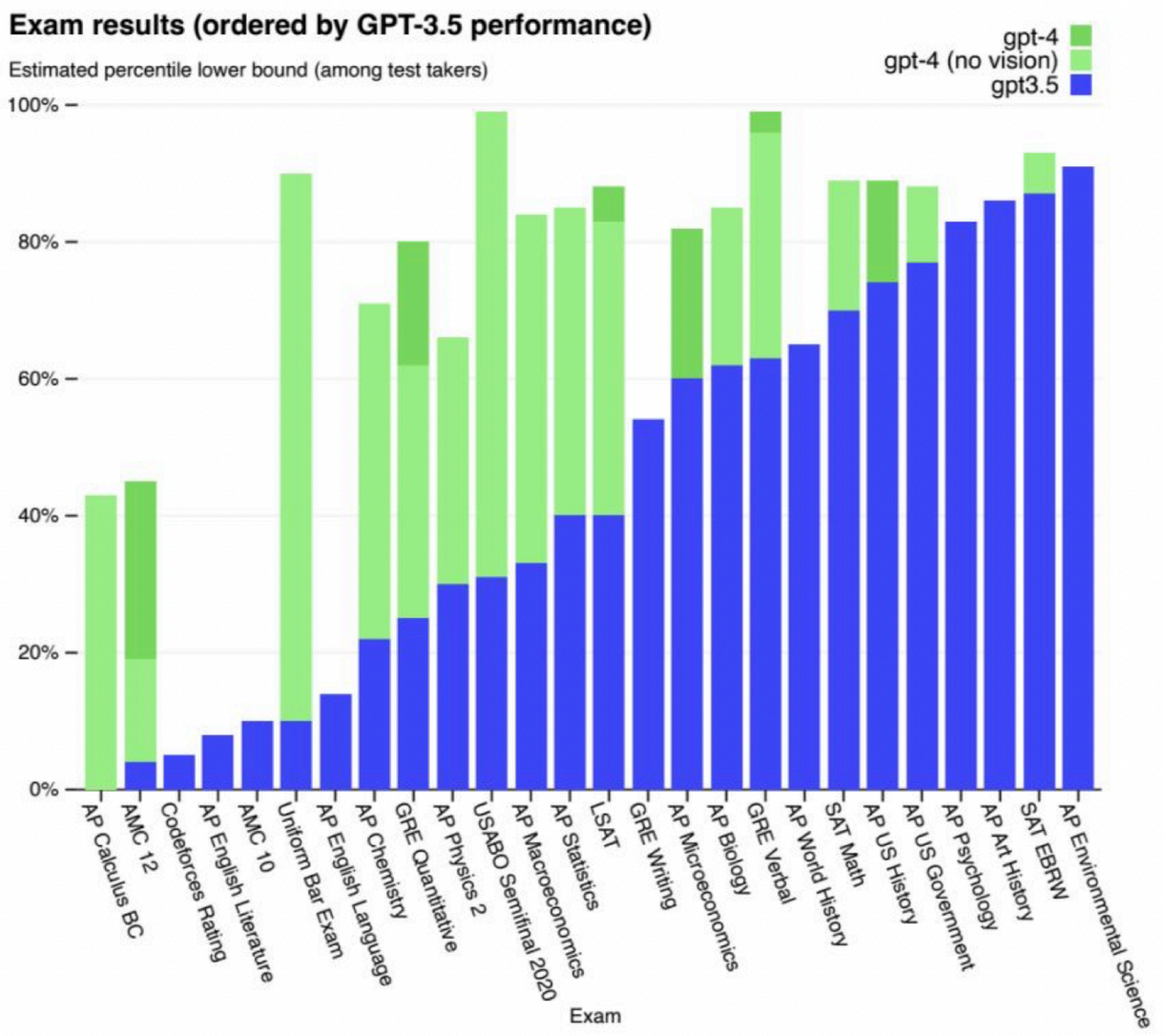

Introducing ChatGPT
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests.
Try ChatGPT ↗
Read about ChatGPT Plus
OpenAI

2024

2024


Amazon-owned
Chinese
Google
Meta / Facebook
Microsoft
OpenAI
Other
BOTS →
billion parameters
BlenderBot1
PLATO-XL
ChatGPT
Bard*
BlenderBot3
BingChat*
Ernie Bot 3.5
Wu Dao 2.0
GLaM
GPT-4*
PanGu-Sigma
Minerva
Titan
PaLM
PaLM2
Gopher
Exaone
SenseChat
Ernie 3.0 Titan
175 Billion
PanGu-Alpha
BLOOM
Ernie Bot
OPT-IML
GPT-3
GLM-130B
WebGPT
LaMDA
FLAN
YaLM 100B
LLaMa2
Chinchilla
LLaMaIbleLM
Falcon LLM
name GPT-J
full owner EleutherAI
bn parameters 6
date Jun 2021
MT5
Dolly 2.0
Alpaca
Sail-7B
GPT-2
BERT
T5
Megatron-11B
Atlas
pre-2020
2020
21
22
2023

2024


Exam results (ordered by GPT-3.5 performance)
Estimated percentile lower bound (among test takers)
gpt-4
gpt-4 (no vision)
gpt3.5
100%
80%
60%
40%
20%
0%
AP Calculus BC
AMC 12
Codeforces Rating
AP English Literature
AMC 10
Uniform Bar Exam
AP English Language
AP Chemistry
GRE Quantitative
AP Physics 2
USABO Semifinal 2020
AP Macroeconomics
AP Statistics
LSAT
GRE Writing
AP Microeconomics
AP Biology
GRE Verbal
AP World History
SAT Math
AP US History
AP US Government
AP Psychology
AP Art History
SAT EBRW
AP Environmental Science
Exam

2025 - …

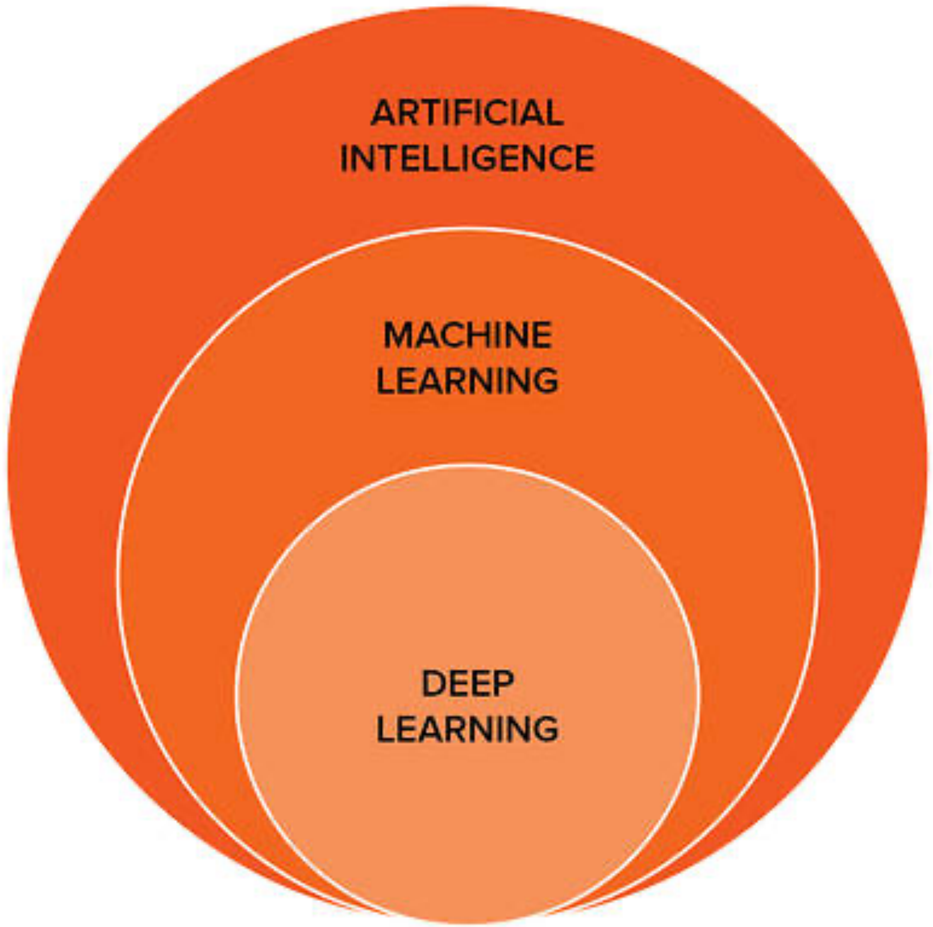
	Computer	Human Brain
Computational Units	8 CPUs, 10^{10} gates	10^{11} neurons
Storage Units	10^{10} bits RAM 10^{13} bits disk	10^{11} neurons 10^{14} synapses
Cycle time	10^{-9} sec	10^{-3} sec
Bandwidth	10^{10} bits/sec	10^{14} bits/sec
Memory updates/sec	10^{10}	10^{14}

Intelligence vs. Rationality

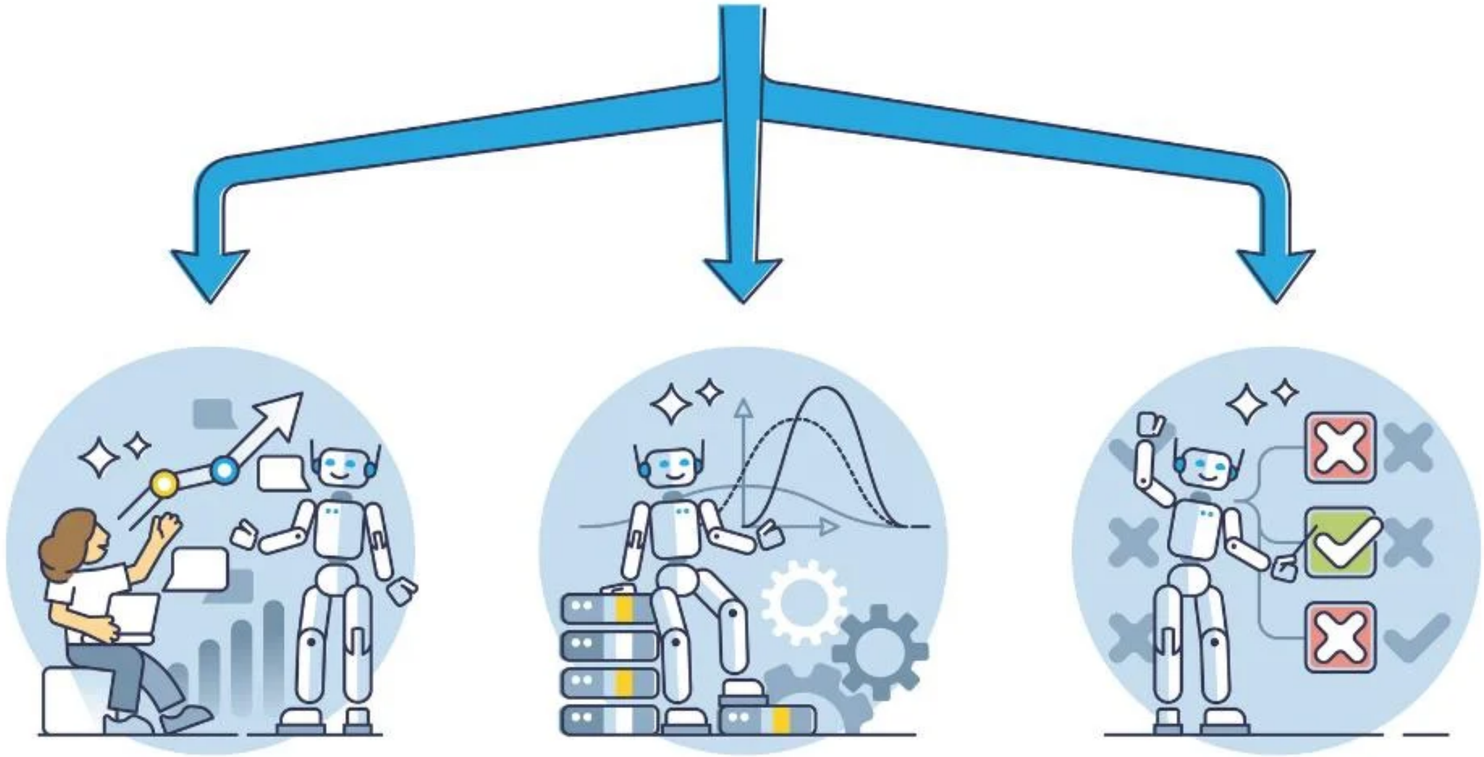
	Like humans	Not necessarily like humans
Think	Systems that think like humans	Systems that think rationally
Act	Systems that act like humans	Systems that act rationally Our focus

Recent Technologies

Machine Learning
Deep Learning



MACHINE LEARNING



SUPERVISED

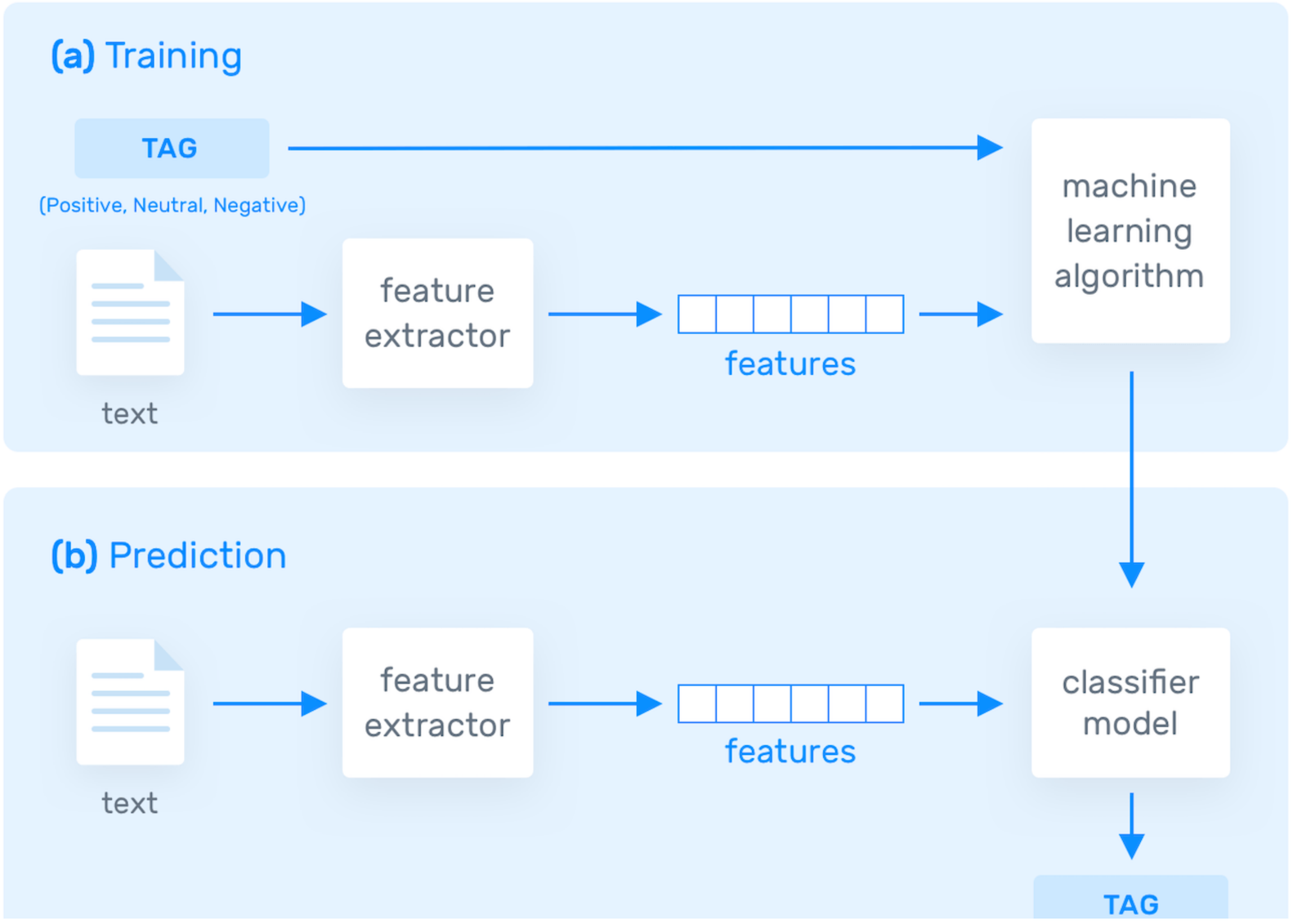
TASK DRIVEN
(PREDICT NEXT VALUE)

UNSUPERVISED

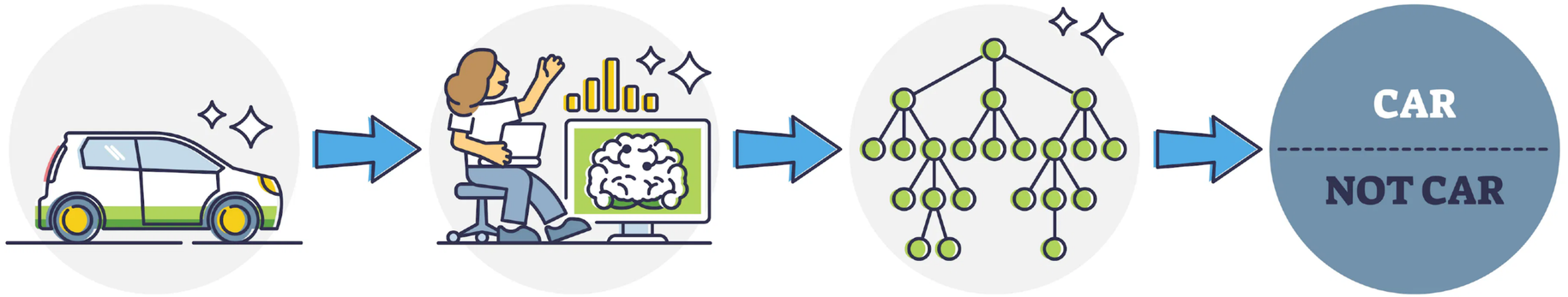
DATA DRIVEN
(IDENTIFY CLUSTERS)

REINFORCEMENT

LEARN
FROM MISTAKES

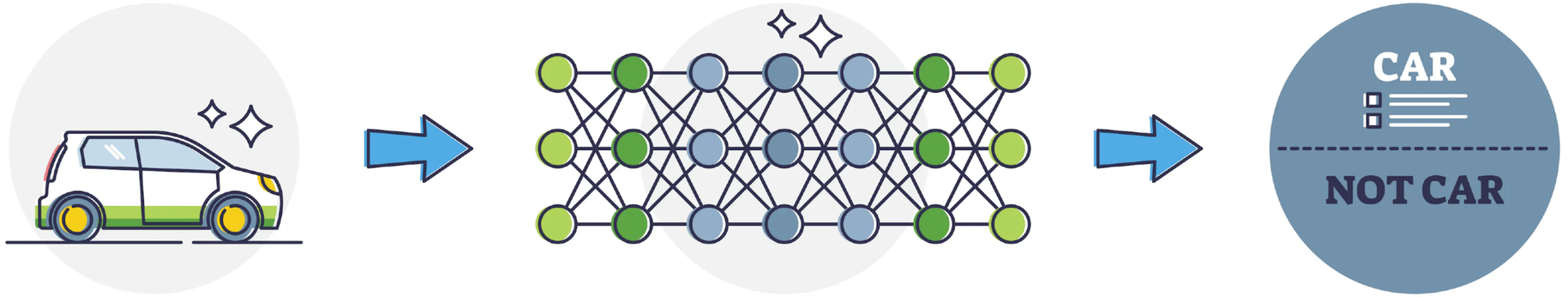

(a) Training
TAG
(Positive, Neutral, Negative)
machine learning algorithm
feature extractor
features
text
(b) Prediction
feature extractor
features
classifier model
text
TAG

MACHINE LEARNING



INPUT | FEATURE EXTRACTION | CLASSIFICATION | OUTPUT

DEEP LEARNING



INPUT | FEATURE EXTRACTION + CLASSIFICATION | OUTPUT

Our journey to Natural Language Processing

(And what you should expect to learn from this course)

Regular Expressions

- woodchuck
- woodchucks
- Woodchuck
- Woodchucks
- Groundhog
- groundhogs

Pattern	Matches
`[wW]oodchuck`	Woodchuck, woodchuck
`[1234567890]`	Any one digit

Regular Expressions

Pattern	Expansion	Matches	Examples
`\d`	`[0-9]`	Any digit	`Fahreneit 451`
`\D`	`[^0-9]`	Any non-digit	`Blue Moon`
`\w`	`[a-ZA-Z0-9_]`	Any alphanumeric or _	`Daiyu`
`\W`	`[^\w]`	Not alphanumeric or _	`Look!`
`\s`	`[ \r\t\n\f]`	Whitespace (space, tab)	`Look␣up`
`\S`	`[^\s]`	Not whitespace	`Look up`

Probabilistic Language Models

- Goal: compute the probability of a sentence or sequence of words:

  $P(W) = P(w_1,w_2,w_3,w_4,w_5...w_n)$

- Related task: probability of an upcoming word:

  $P(w_5|w_1,w_2,w_3,w_4)$

- A model that computes either of these:

  $P(W)$ or $P(w_n|w_1,w_2...w_{n-1})$ is called a **language model**.

Probabilistic Language Models

- The Chain Rule in General

$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$

P("its water is so transparent") =

$P(its) \times P(water|its) \times P(is|its\ water)$

$\times P(so|its\ water\ is) \times P(transparent|its\ water\ is\ so)$

N-grams

P(<s> I want english food </s>) =

P(I|<s>)

× P(want|I)

× P(english|want)

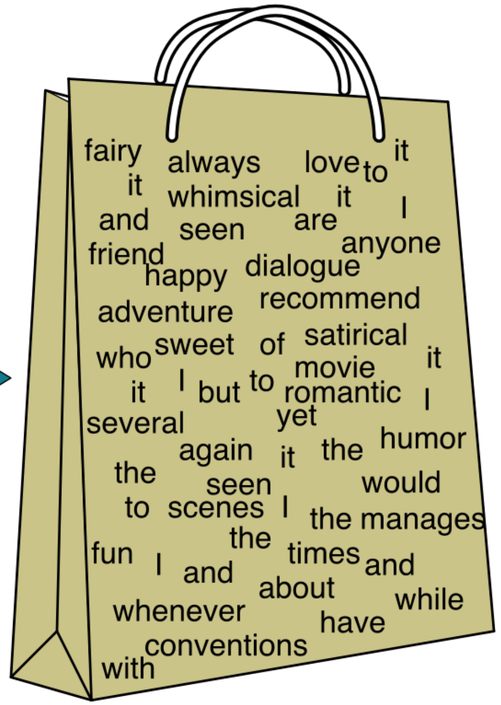
× P(food|english)

× P(</s>|food)

= .000031

Bag-of-Words Model

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Bag-of-Words Model: Word Vectors

	As You Like It	Twelfth Night	Julius Caesar	Henry V
battle	1	0	7	13
good	114	80	62	89
fool	36	58	1	4
wit	20	15	2	3

Word Embeddings

…lemon, a [tablespoon of apricot jam, a] pinch…

c1 c2 [target] c3 c4

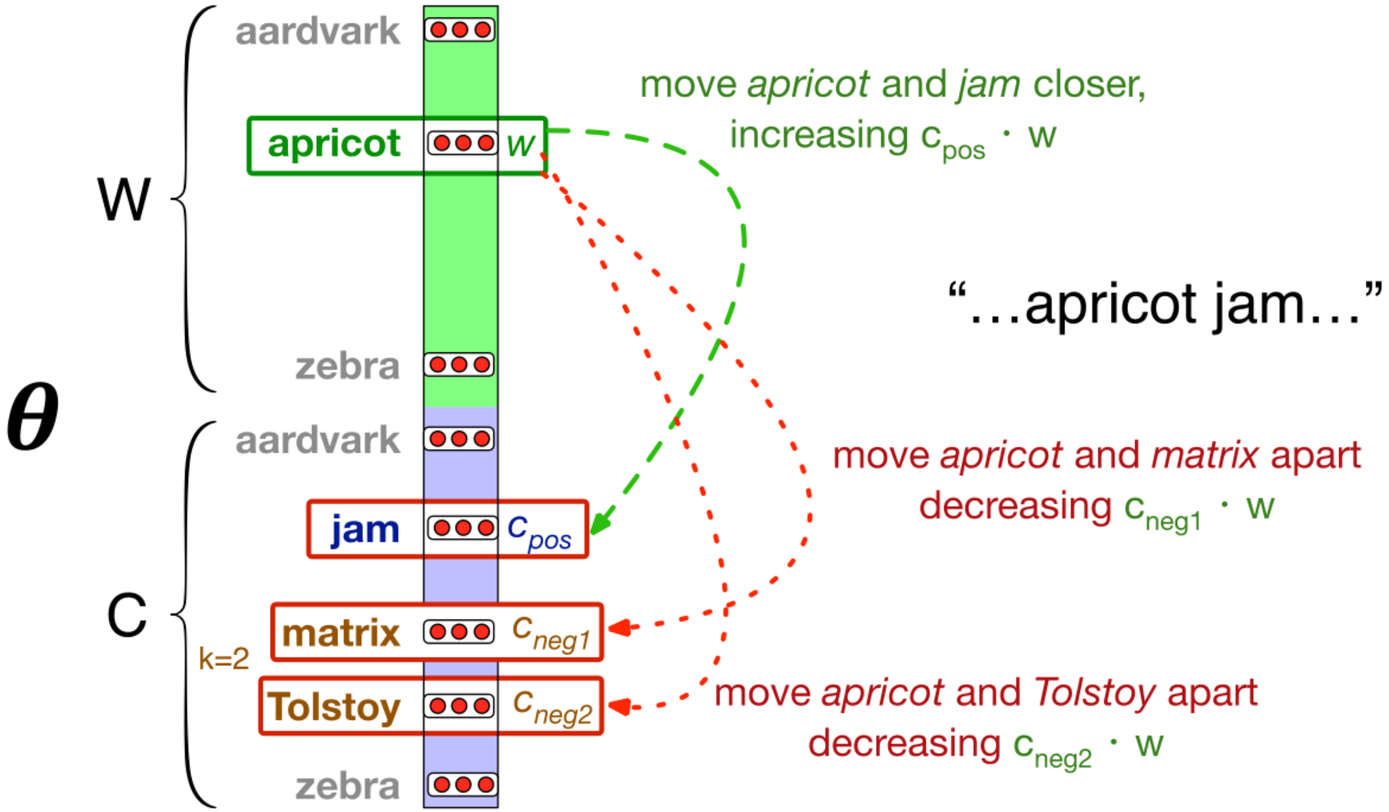
positive examples +

t	c
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

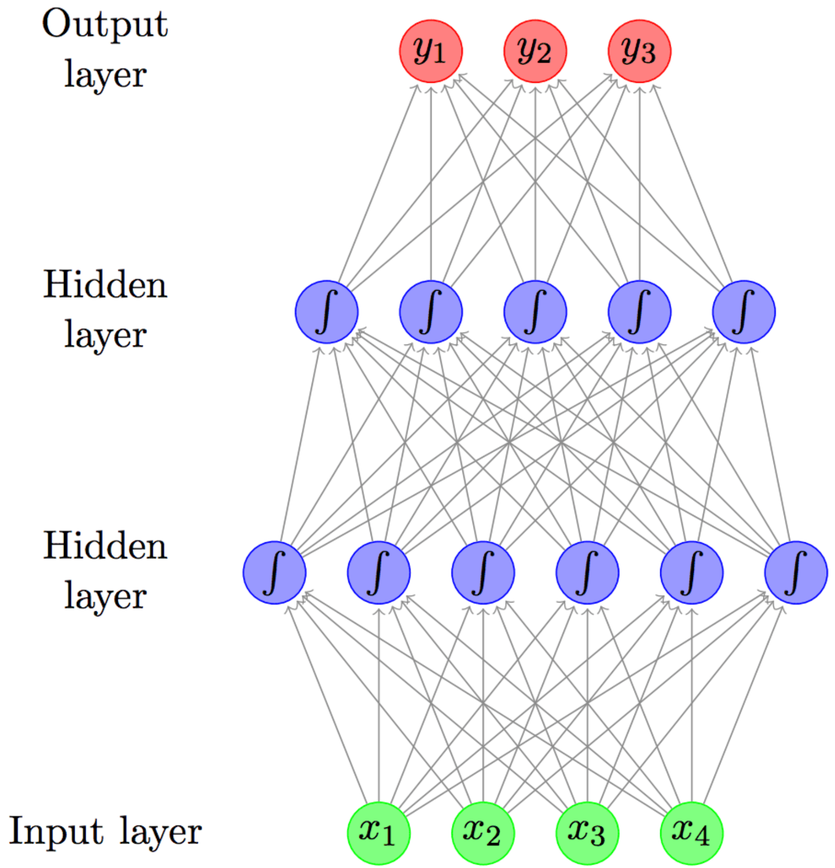
negative examples -

t	c	t	c
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

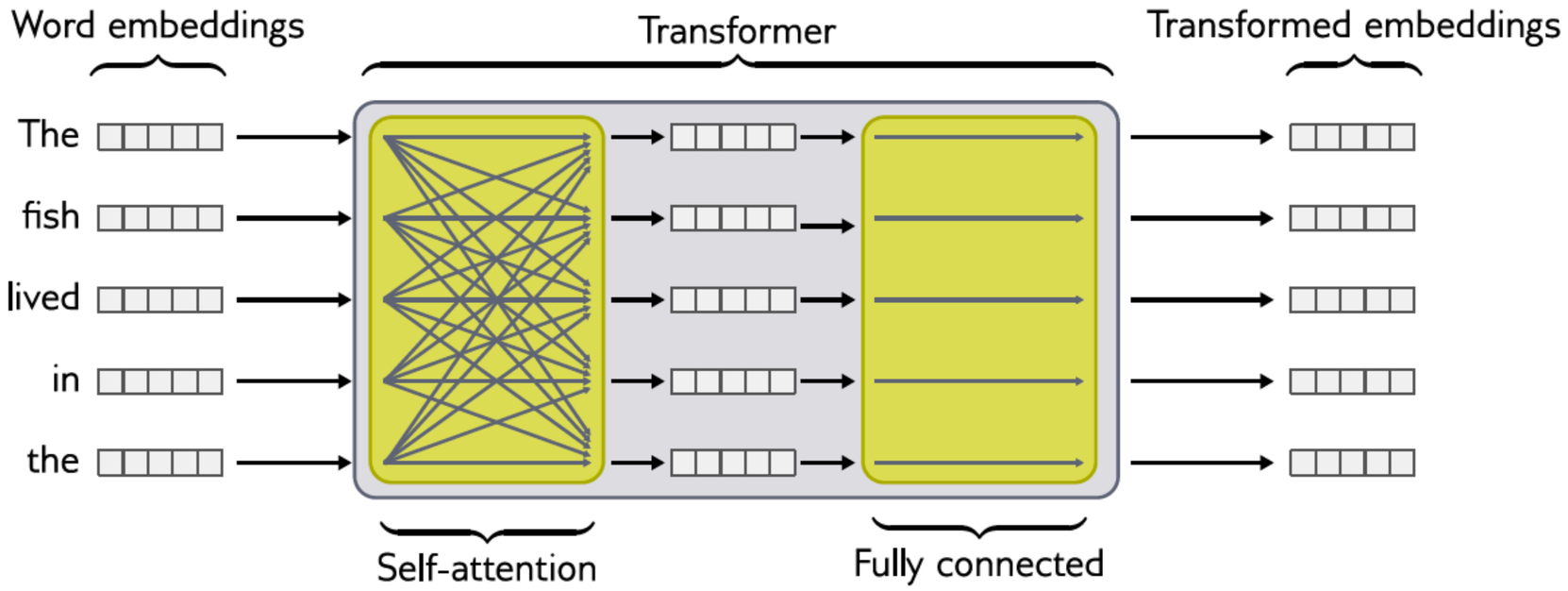
Word Embeddings



Neural Models



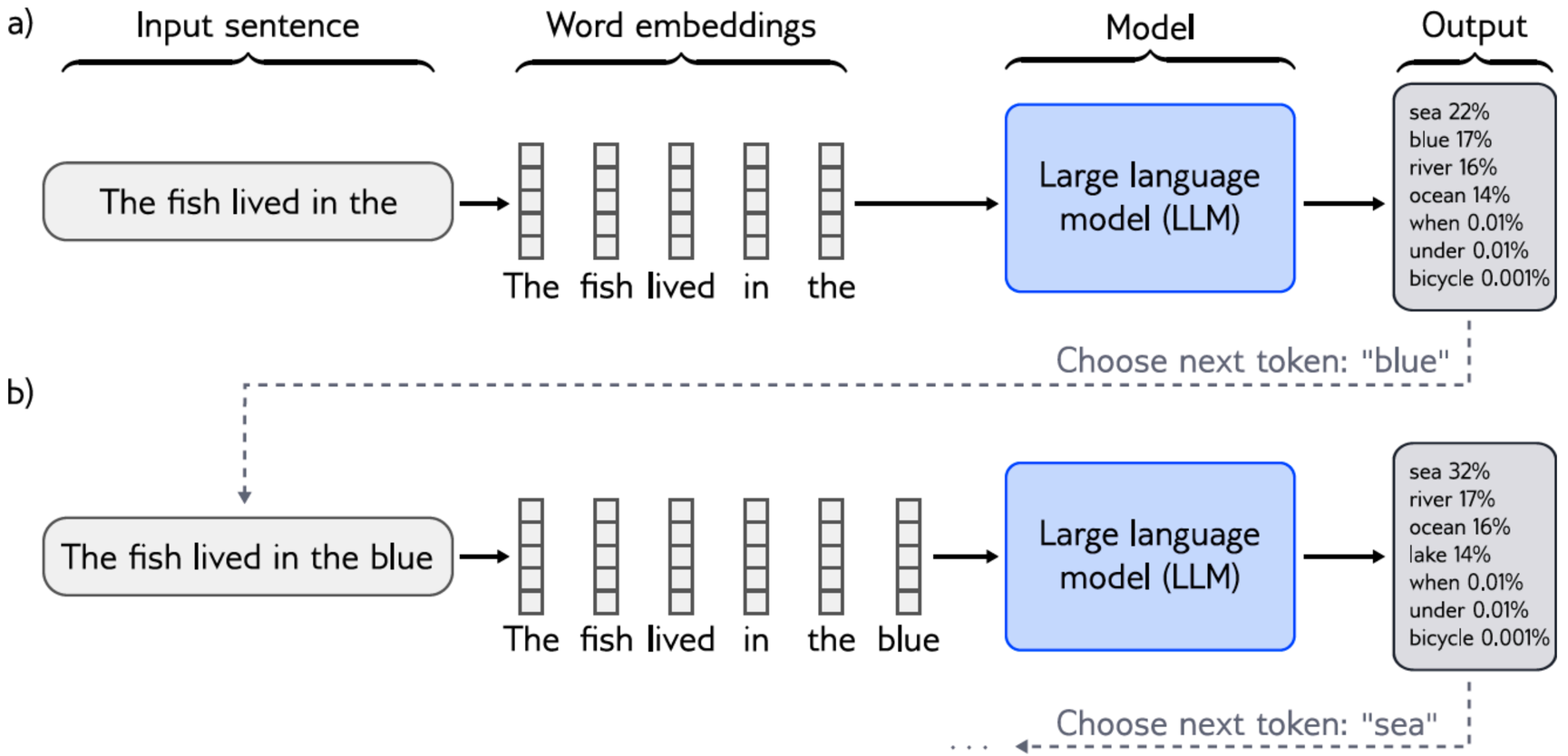
Transformer Language Models



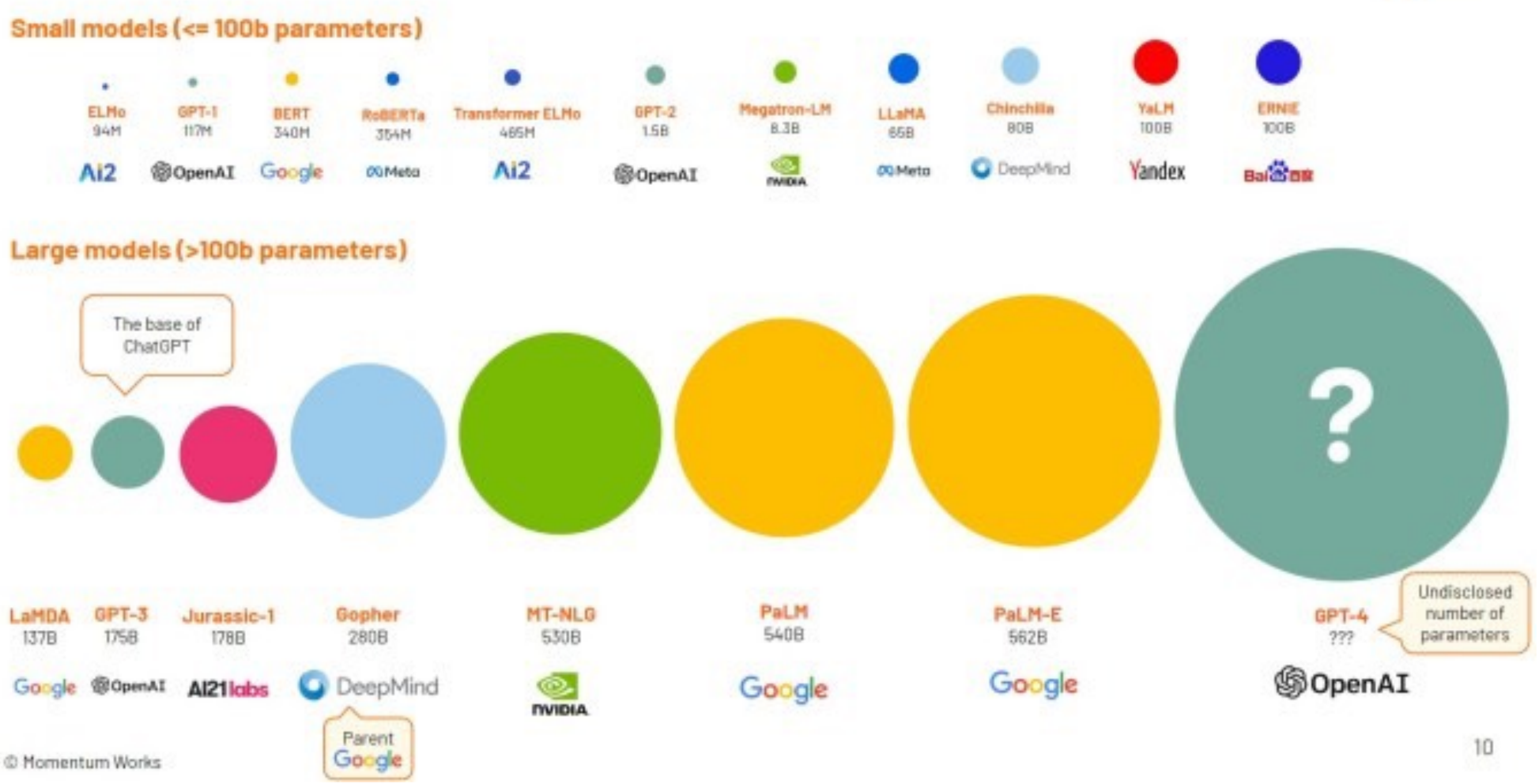
Transformer Language Models

- Encoder (BERT)
- Decoder (GPT3)
- Encoder-decoder (Translation)

Transformer Language Models: GPT



Large Language Models (LLMs)



Large Language Models (LLMs)

Prompting: Prompting a language model w/o training

If the following sentences is about “sports” reply “sports”. Otherwise reply “other”.

{X}

→ LM

Fine-tuning: Machine learning from paired data <*X*, *Y*>

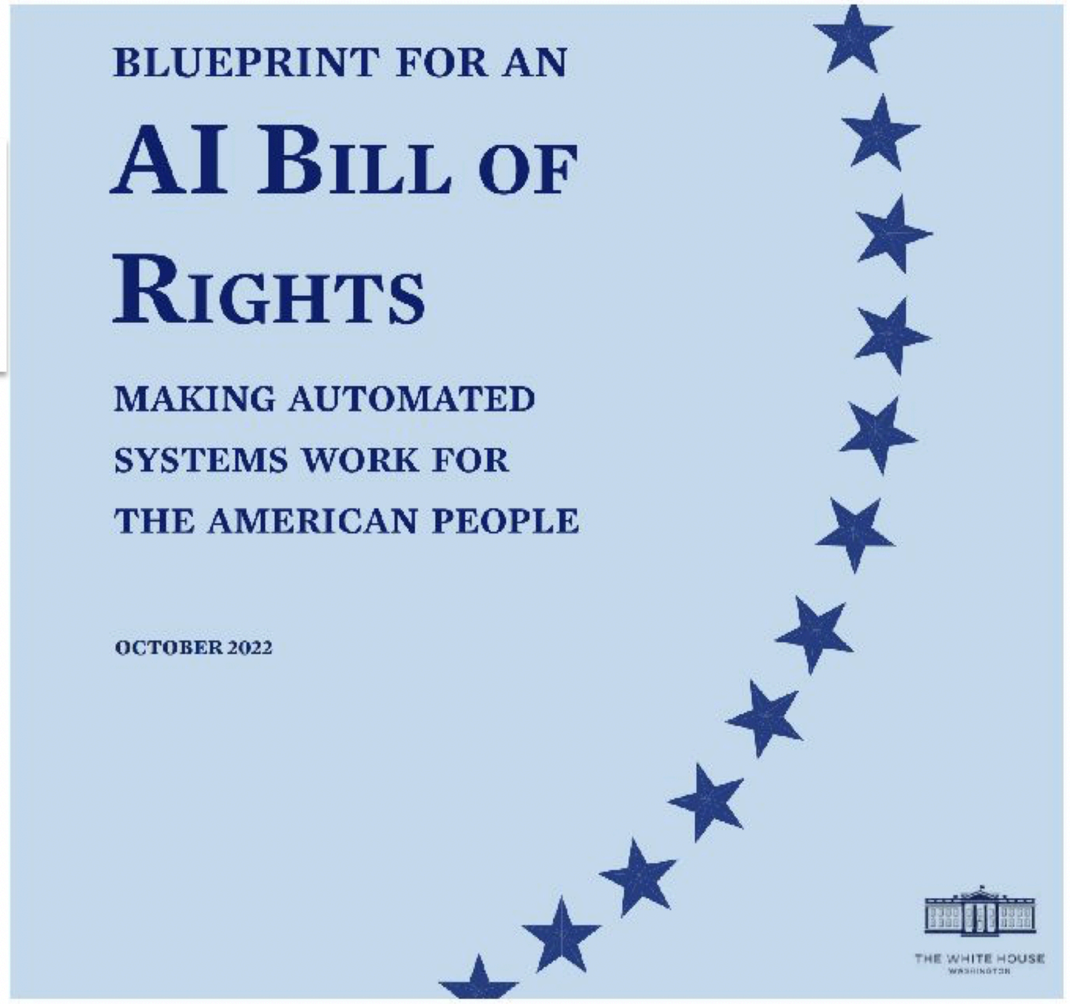
I love to play baseball.	sports
The stock price is going up.	other
He got a hat-trick yesterday.	sports
He is wearing tennis shoes.	other

→ Training → Model

Large Language Models (LLMs)

Importance of Responsible AI

- Robustness: Safe and Effective Systems
- Fairness: Algorithmic Discrimination Protections
- Data Privacy
- Notice and Explanation
- Human Alternatives, Consideration, and Fallback



Large Language Models (LLMs)

- Who controls AI? Centralized vs. decentralized control; open vs. closed source

- Trustworthiness

- Robustness: Test-time attacks vs. training-time attacks

- Privacy

- Fairness

- Toxicity

- AI Safety: Misuse/abuse of AI – Super intelligence

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

Thanks for your participation!

Çağrı Toraman
30.09.2025