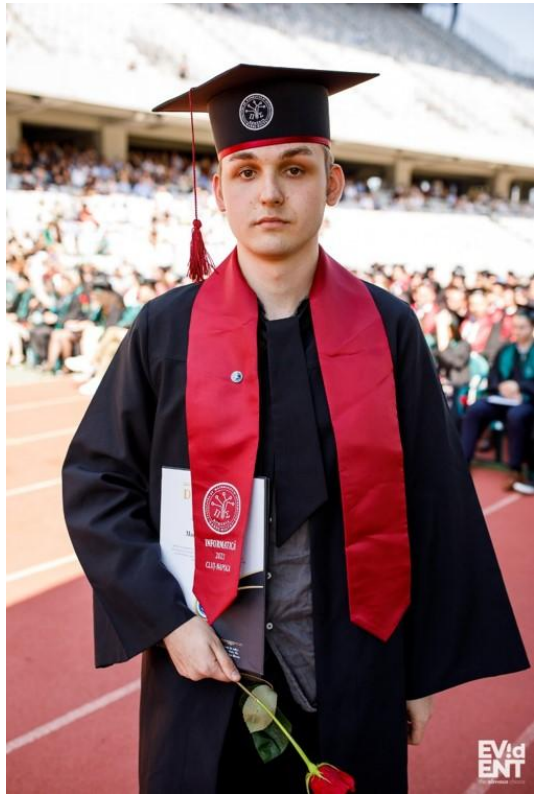
Inteligență Artificială Generativă - LLMs

AI Curs 7 – 09.04.2025

Capitole

- Introducere în Generative AI – Alexandru Manole
- Embeddings – Răzvan Petec
- Stable Diffusion?

Despre mine



- Student la doctorat în anul II la UBB FMI
- Domenii de interes: Computer Vision, Multitask Models, Image Generation
- alexandru.manole@ubbcluj.ro

Despre voi

- https://www.menti.com/
- cod: **5193 0229**

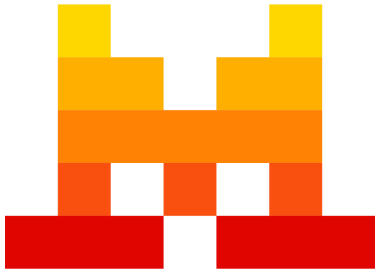
Introducere în Generative AI

ChatGPT

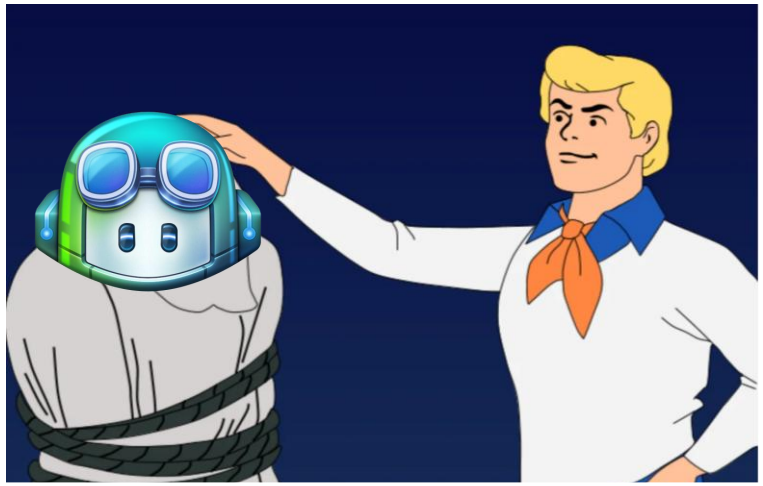
Gemini

Claude

DALL·E

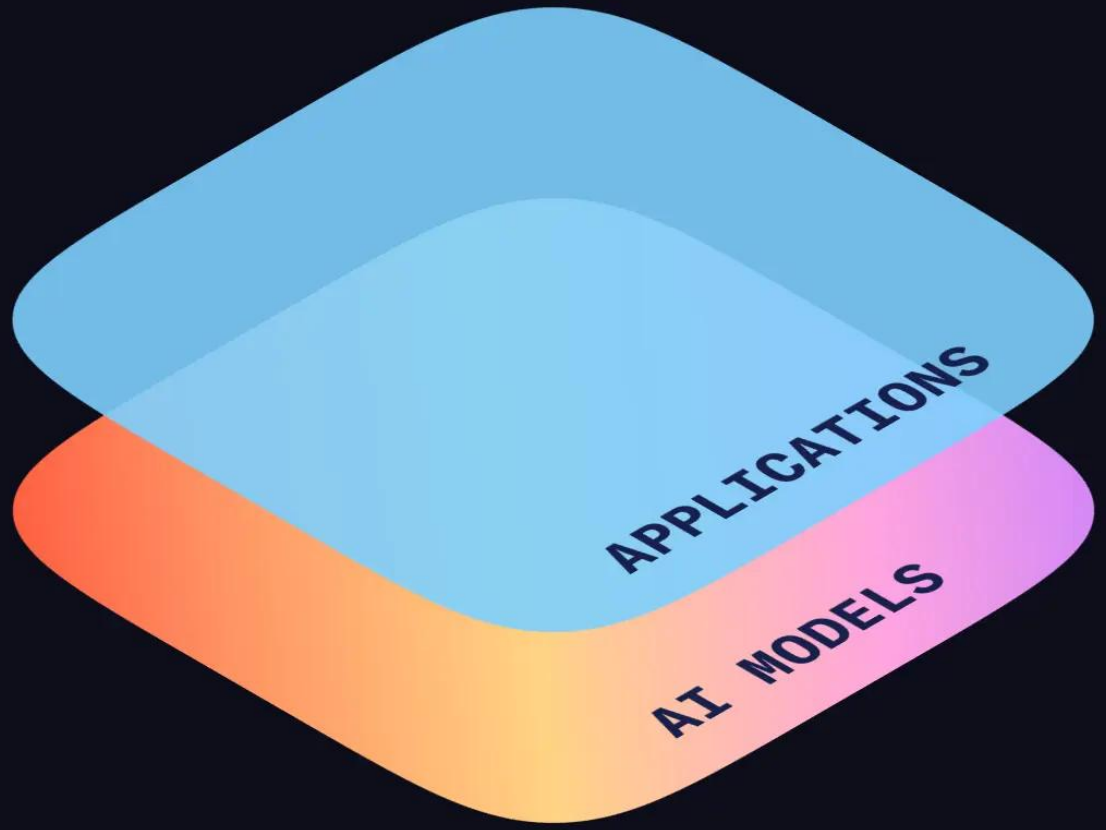


Introducere în Generative AI



- ChatGPT, Gemini, Copilot, Claude sunt sisteme software / tool-uri complexe
- Aceste au în spate modele inteligente generative (LLMs)

The Generative Tech Stack



TOOLS THAT COLLABORATE WITH AI MODELS.

WORKFLOWS, SECURITY, NETWORK EFFECTS, PAYMENTS, ETC. (10,000'S OF THESE)

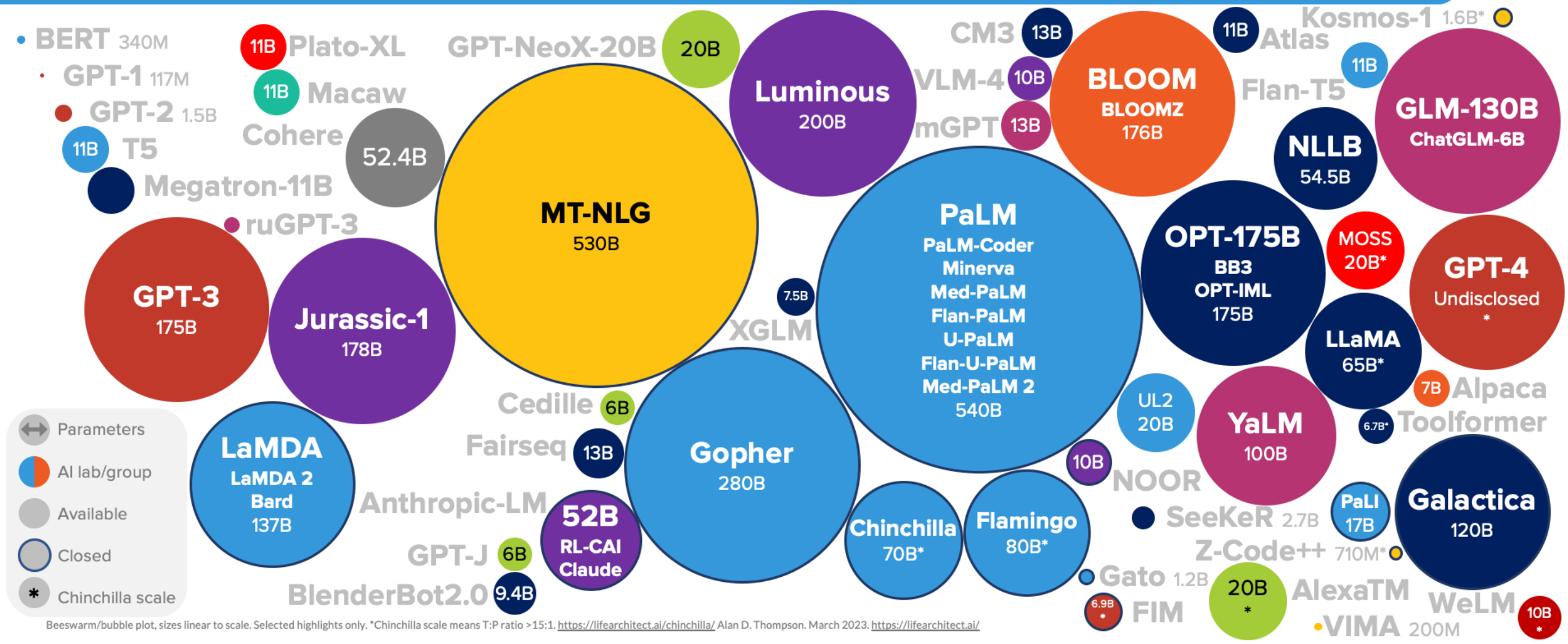
MODELS GENERATE UNIQUE AND NOVEL OUTPUT.

GPT-3, STABLE DIFFUSION, CUSTOM DATA SETS, ETC. (1,000'S OF THESE)

SOURCE: NFX.COM (2022)

NfX

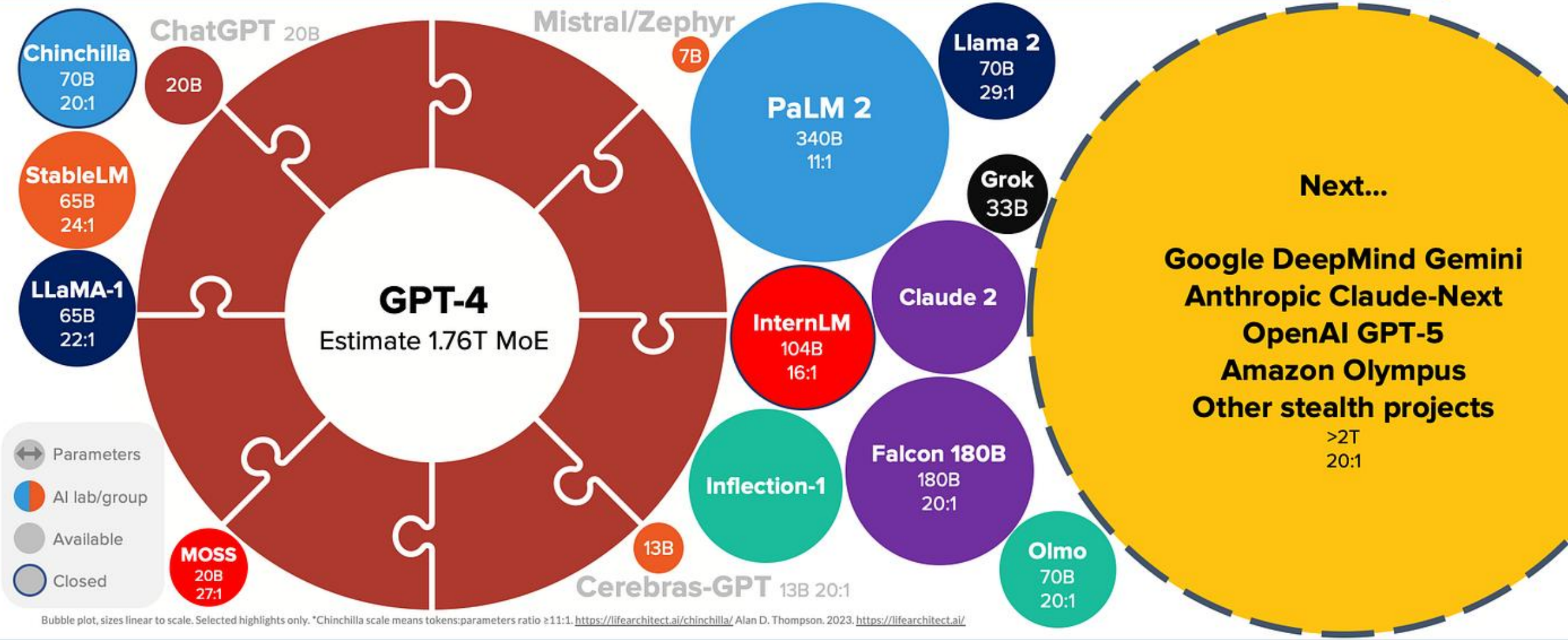
LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

2023-2024 OPTIMAL LANGUAGE MODELS

NOV/ 2023



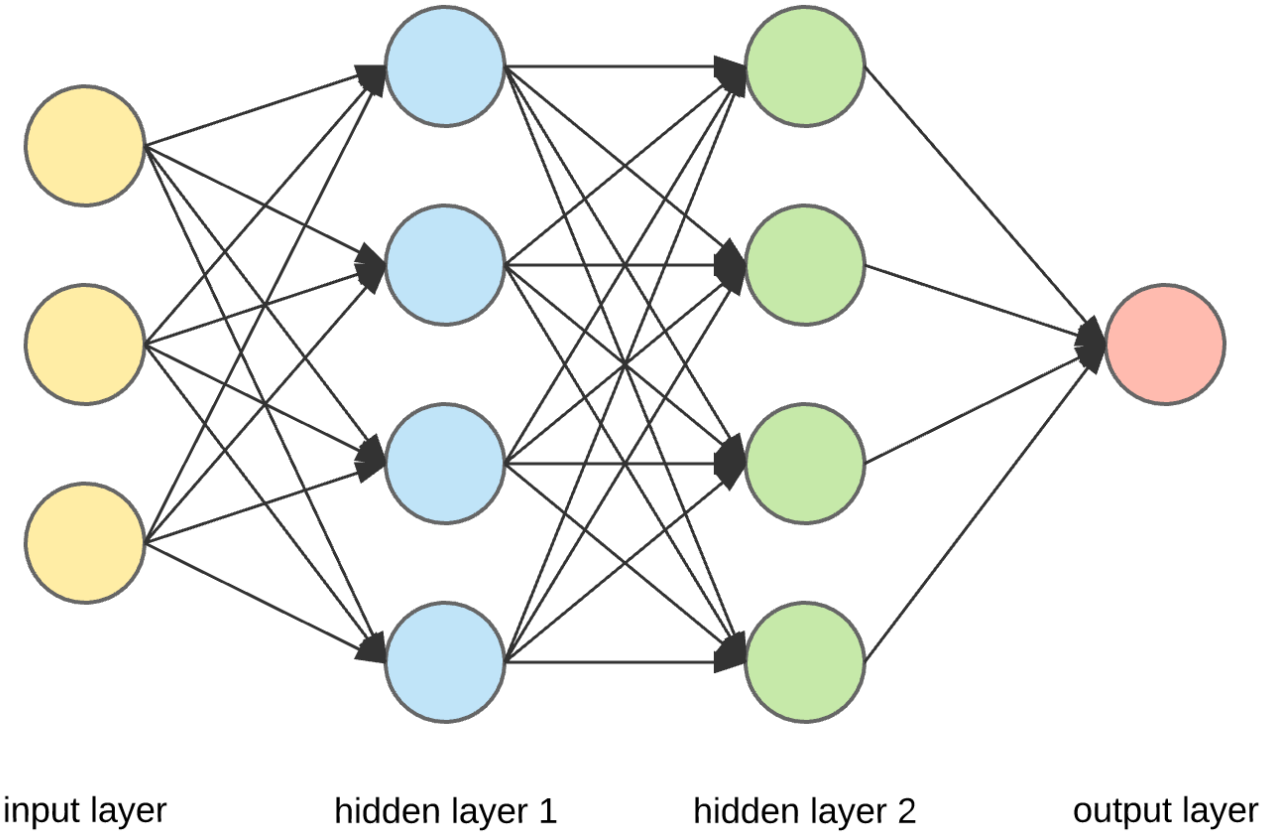
Bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means tokens:parameters ratio ≥11:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. 2023. https://lifearchitect.ai/

Model	Lab	Playground	Parameters (B)	Tokens trained (B)	Ratio Tokens:Params (Chinchilla scaling≥20:1)	ALScore "ALScore" i Sqr Root of	MMLU	MMLU -Pro	GPQA	HLE	Training dataset	Announced ▼	Public?	Paper / Repo	Arch	Tags
AuroraGPT (ScienceGF	Argonne National L	https://lifearchite	*2000*	*30000*	15:1							*TBA*				
DeepSeek-R2	DeepSeek-AI	https://www.reut										*TBA*			*MoE*	*Reasoning* *SOTA*
ERNIE 5	Baidu	https://lifearchite										*TBA*				
Grok-4	xAI	https://lifearchite										*TBA*			*MoE*	*Reasoning* *SOTA*
GPT-5	OpenAI	https://lifearchite	*5400*	*114000*	22:1							*TBA*			*MoE*	*SOTA*
GPT-6	OpenAI	https://lifearchite										*TBA*				*SOTA*
Llama 4 Reasoning	Meta AI	https://ai.meta.cc										*TBA*		https://ai.m	MoE	SOTA Reasoning
~~*MAI-1*~~	~~Microsoft~~	~~https://arstechni~~	~~500~~	~~*10000*~~	~~20:1~~	~~7.5~~						~~*TBA*~~		~~https://ww~~	~~*Dense*~~	
o4	OpenAI	https://lifearchite										*TBA*				*Reasoning* *SOTA*
Llama 4 Behemoth	Meta AI	https://ai.meta.cc	2000	30000	15:1	25.8		82.2	73.7			Apr/2025		https://ai.m	MoE	SOTA
Llama 4 Maverick	Meta AI	https://ai.meta.cc	400	22000	55:1	9.9		80.5	69.8			Apr/2025		https://ai.m	MoE	SOTA
Qwerky-72B	Featherless AI	https://featherles	72	18000	250:1	3.8	77.46				synthetic, web-scale	Apr/2025		https://featl	Dense	
Agentic-Tx	Google DeepMind	https://github.co	*200*	*20000*	100:1	6.7			62.4	14.5	synthetic, web-scale	Mar/2025		https://stor	*MoE*	
TxGemma	Google DeepMind	https://huggingfa	27	14000	519:1	2.0					synthetic, web-scale	Mar/2025		https://stor	*MoE*	Reasoning
Gemini 2.5 Pro	Google DeepMind	https://aistudio.g	*200*	*20000*	100:1	6.7			84	18.8	synthetic, web-scale	Mar/2025		https://blog	*MoE*	Reasoning SOTA
DeepSeek-V3 0324	DeepSeek-AI	https://chat.deep	685	14800	22:1	10.6		81.2	68.4		synthetic, web-scale	Mar/2025		https://hugg	MoE	SOTA
Llama-3.3-Nemotron-	NVIDIA	https://huggingfa	49	*15040*	307:1	2.9			66.67		web-scale	Mar/2025		https://buil	Dense	Reasoning
EXAONE Deep	LG	https://huggingfa	32	6500	204:1	1.5	83	74	66.1		web-scale	Mar/2025		https://arxi	Dense	Reasoning
Mistral Small 3.1	Mistral	https://huggingfa	24	*8000*	334:1	1.5	81.01	56.03	37.5		web-scale	Mar/2025		https://mist	Dense	
ERNIE 4.5	Baidu	https://yiyan.bai						*79*	*55*		synthetic, web-scale	Mar/2025		https://ww	*MoE*	
X1	Baidu	https://yiyan.bai									synthetic, web-scale	Mar/2025		https://ww	*MoE*	Reasoning
OLMo 2 32B	Allen AI	https://playgrour	32	6400	200:1	1.5	78				synthetic, web-scale	Mar/2025		https://alle	Dense	
Command A	Cohere	https://dashboar	111	*8000*	73:1	3.1	85				synthetic, web-scale	Mar/2025		https://hugg	Dense	
Gemini Robotics	Google DeepMind		*200*	*20000*	100:1	6.7		*79.1*	*64.7*		synthetic, web-scale	Mar/2025		https://stor	*MoE*	

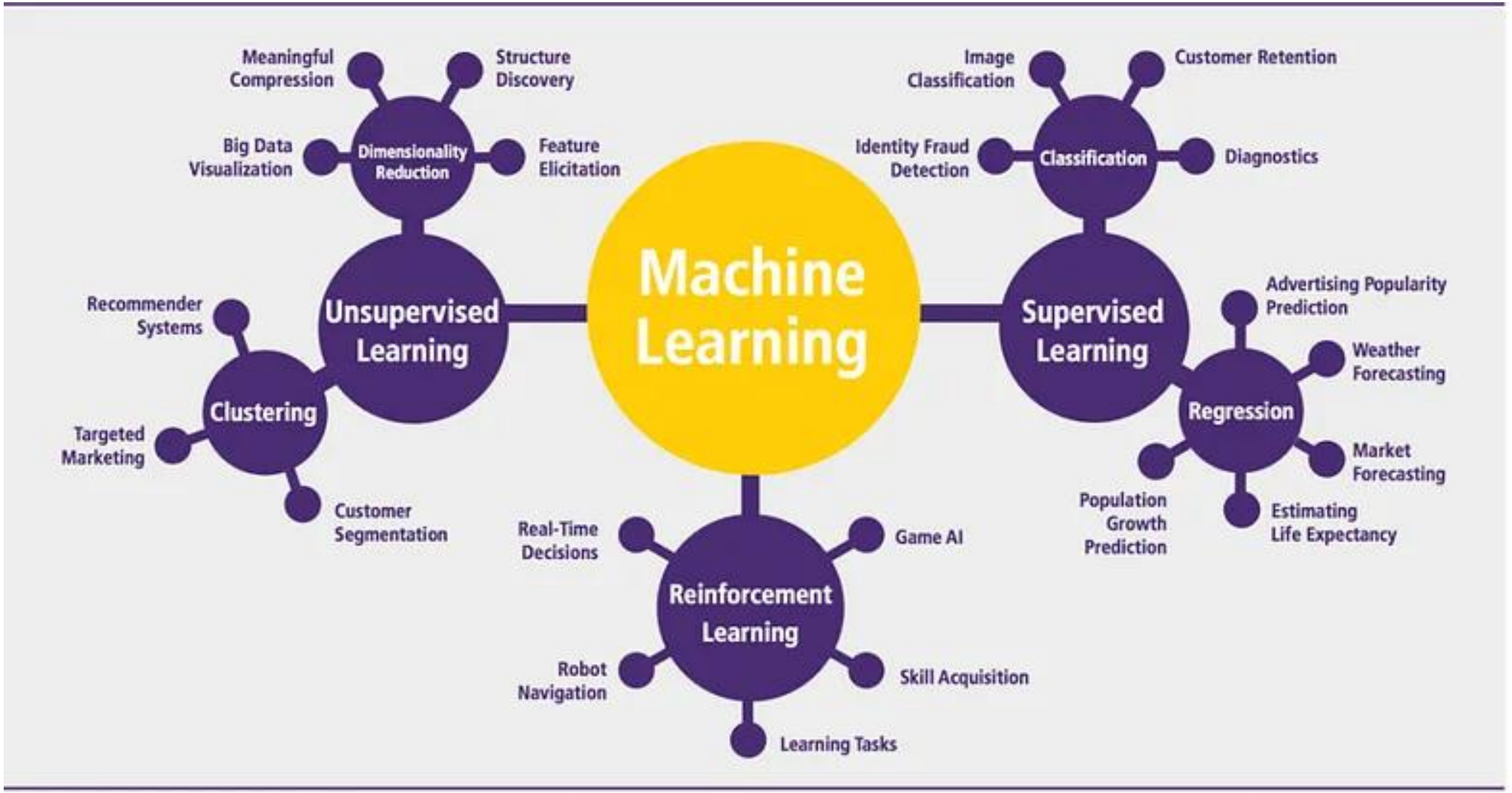
https://lifearchitect.ai/

Elemente AI / Machine Learning:

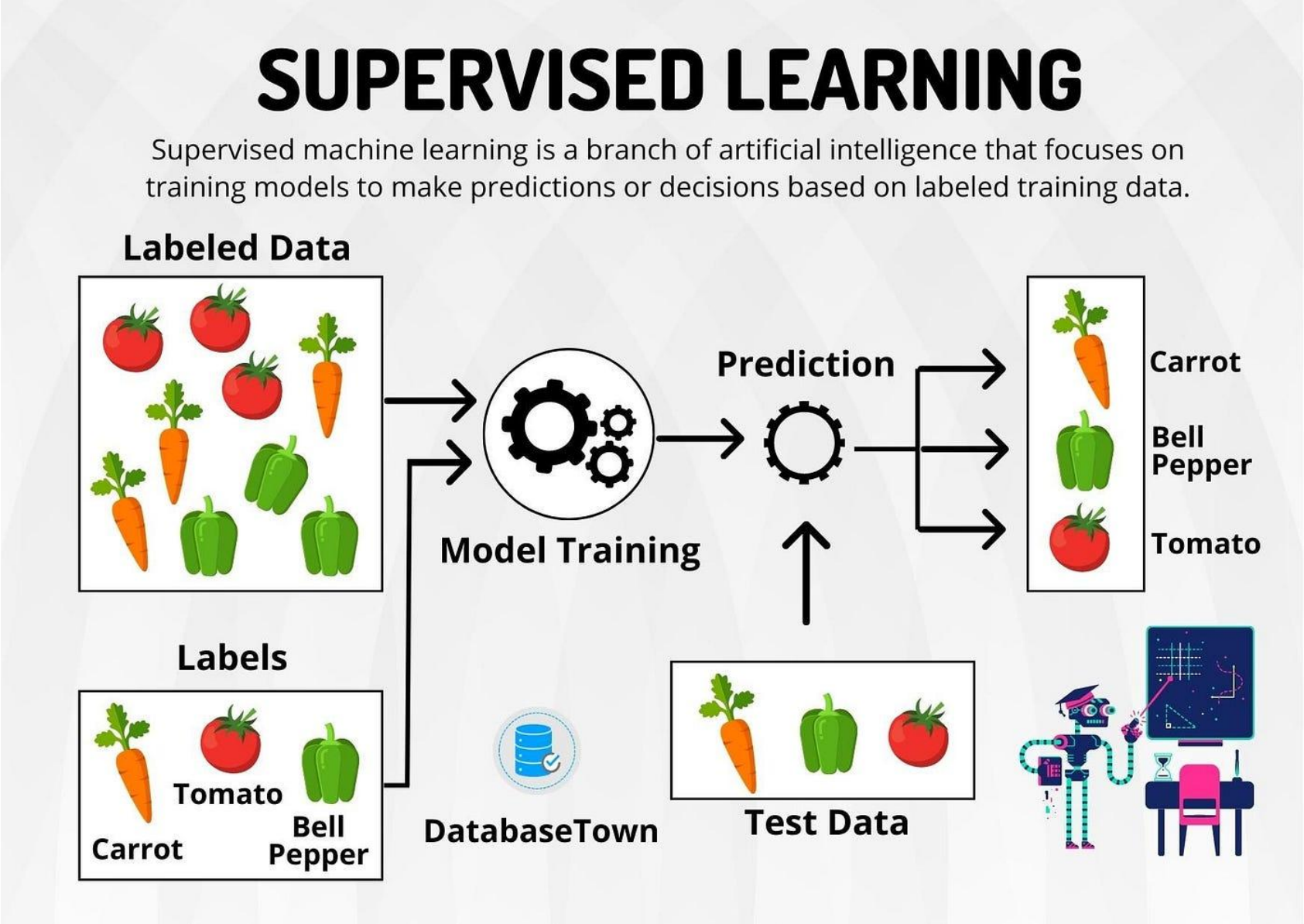
1. Algoritm / Model inteligent
2. Date
   - Date de intrare
   - Date de ieșire



Învățare automată

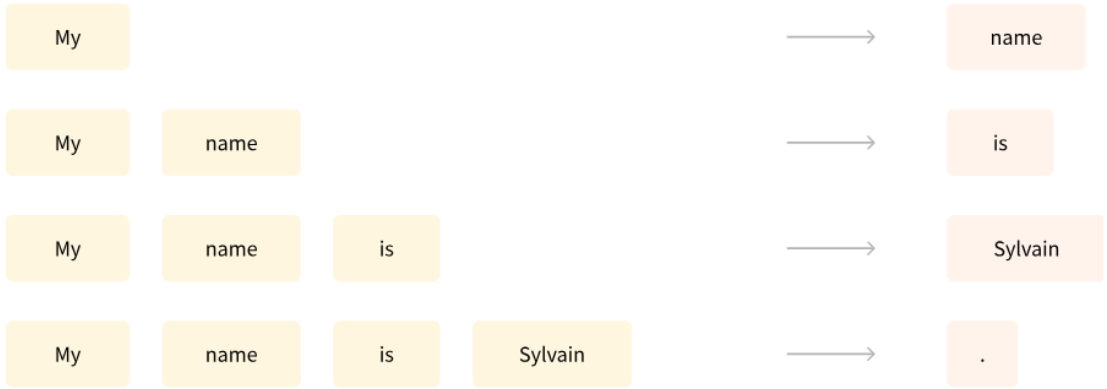


Învățare supervizată

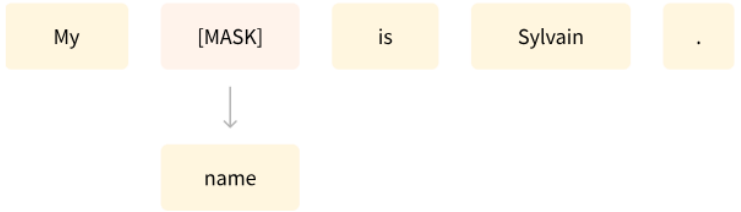


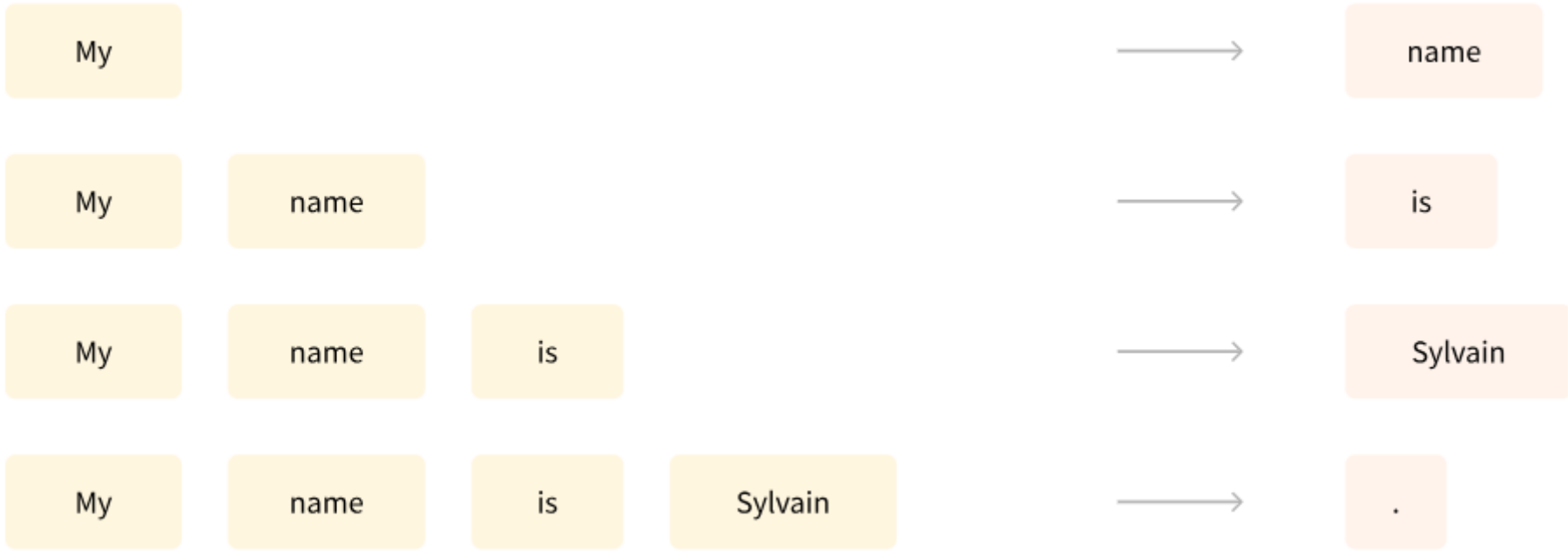
Ce clasifică LLM-urile?

Ce clasifică LLM-urile?

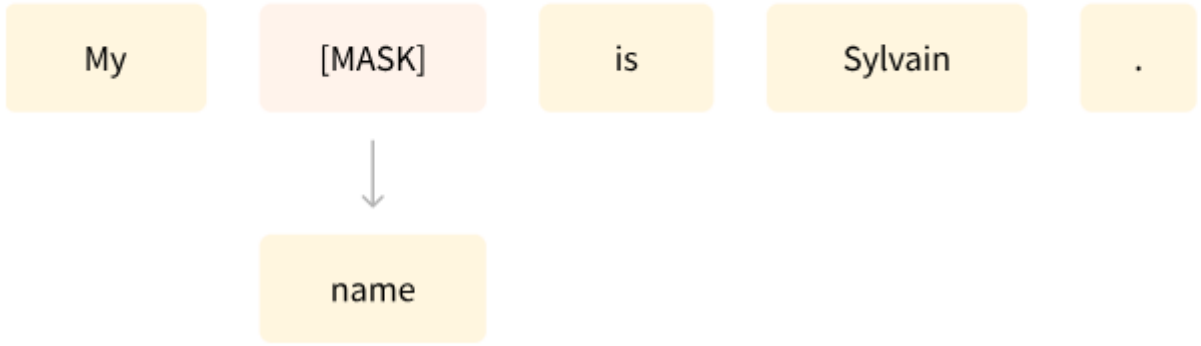


Another example is *masked language modeling*, in which the model predicts a masked word in the sentence.





Another example is *masked language modeling*, in which the model predicts a masked word in the sentence.

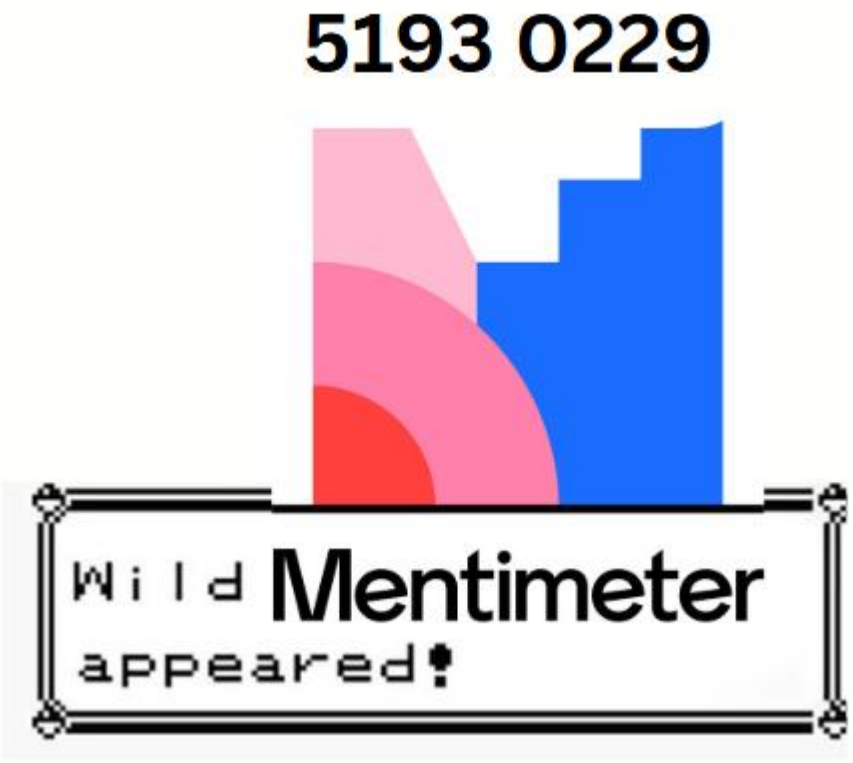


Proces de antrenare

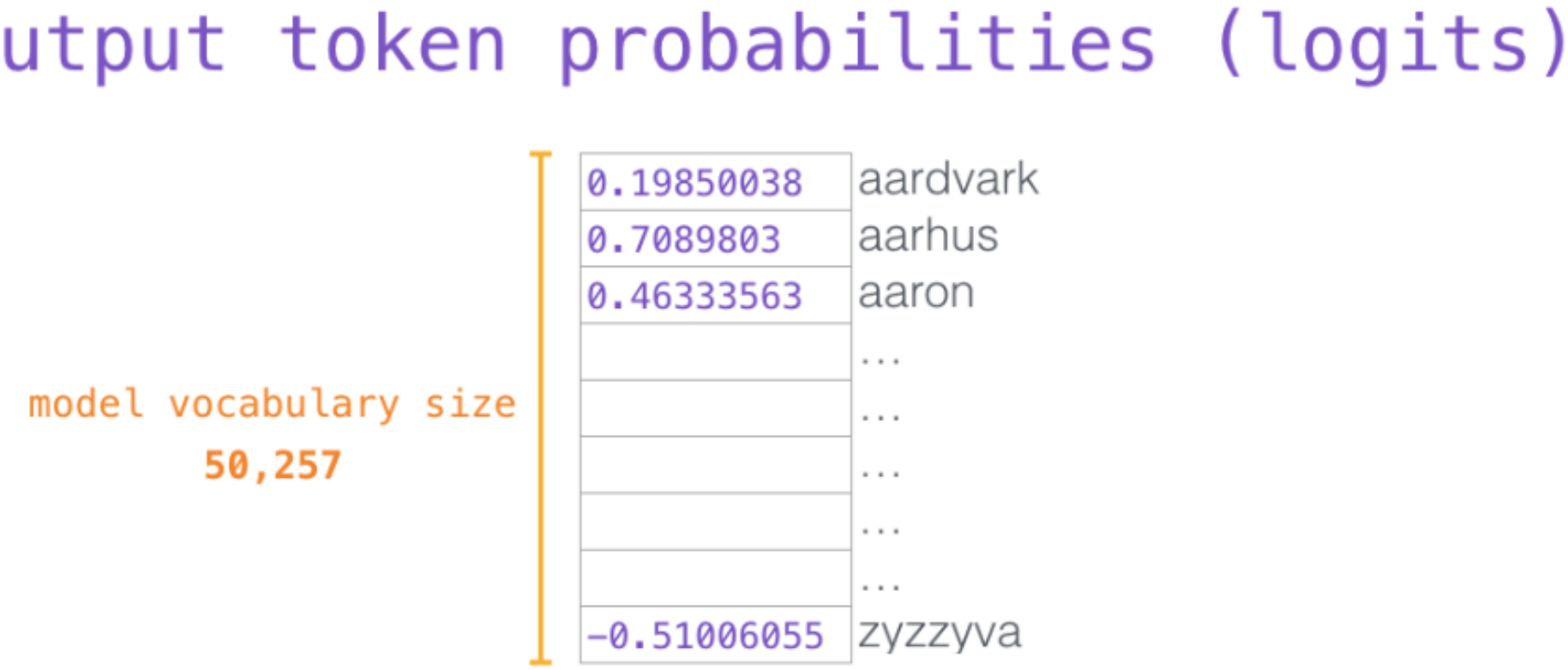
Modelele gene

cuv 1 cuv

- Pornind de la un șir
  cuvine modelul
  zice ce cuvânt ar
  ea urma



Rezultatul predicției



Cuvinte în limba română?

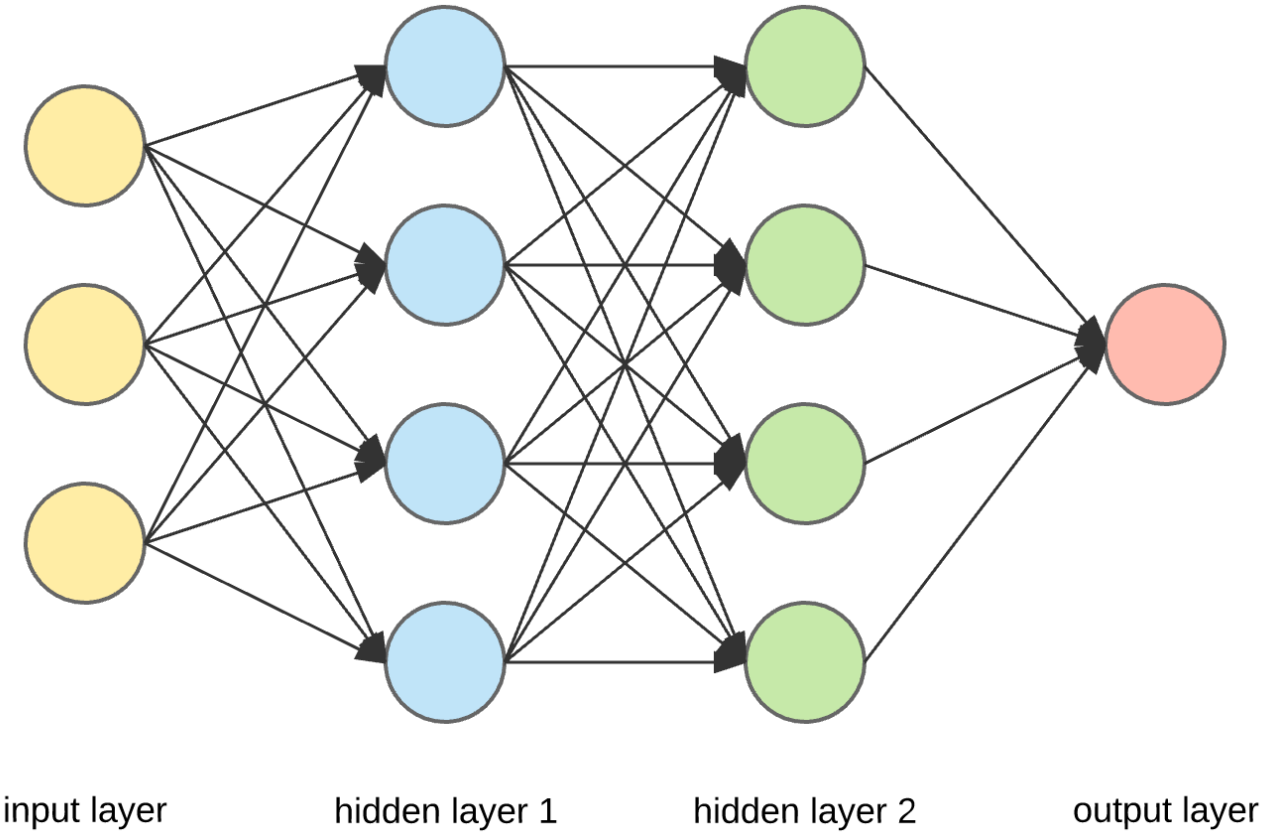
- https://www.menti.com/
- cod: **5193 0229**

Cuvinte în limba română

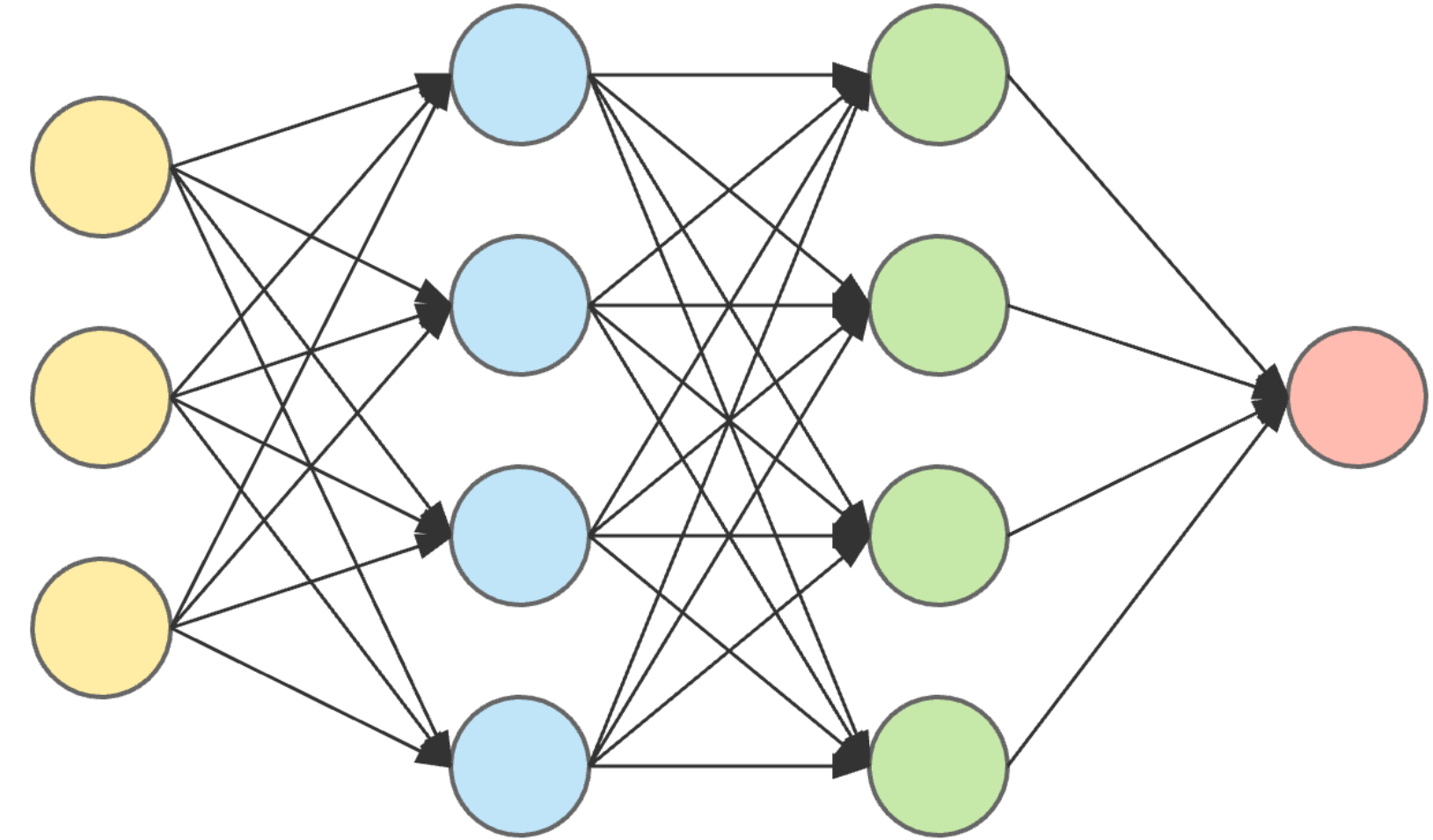
- DOOM – “peste 62.000 cuvinte”
- Dexonline.ro – 75.399 cuvinte
- DLR – 175.000 cuvinte

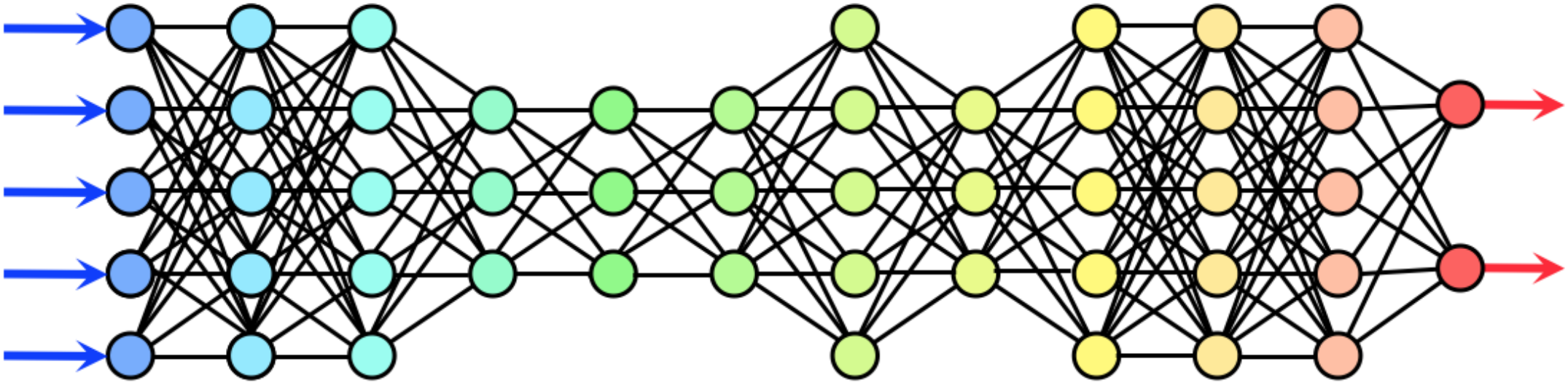
Elemente AI / Machine Learning:

1. Algoritm / Model inteligent
2. Date
   - Date de intrare
   - Date de ieșire



Cum se numește acest algoritm inteligent:
menti.com cod: **5193 0229**



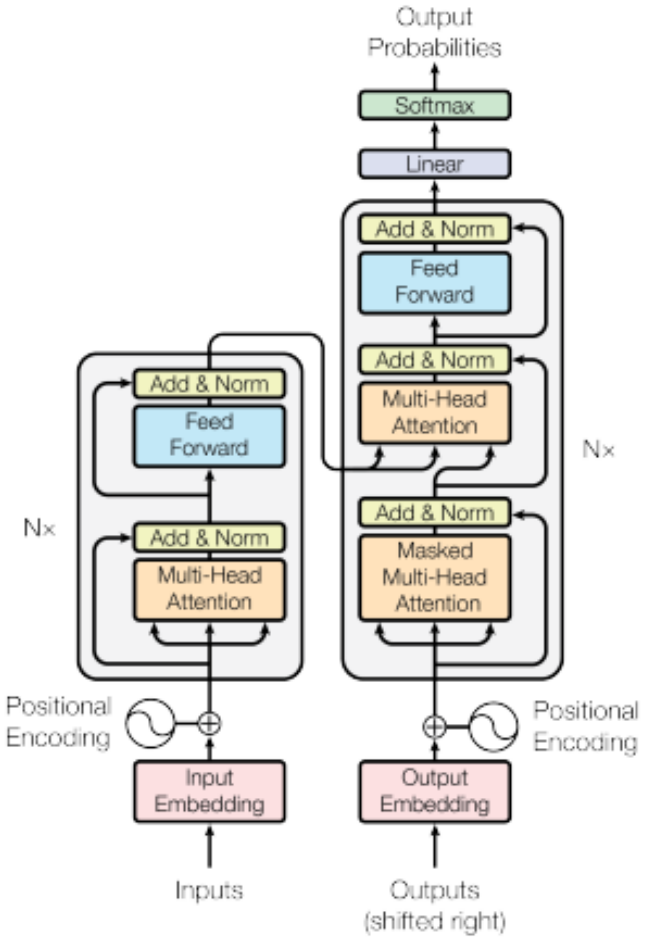


Input Layer

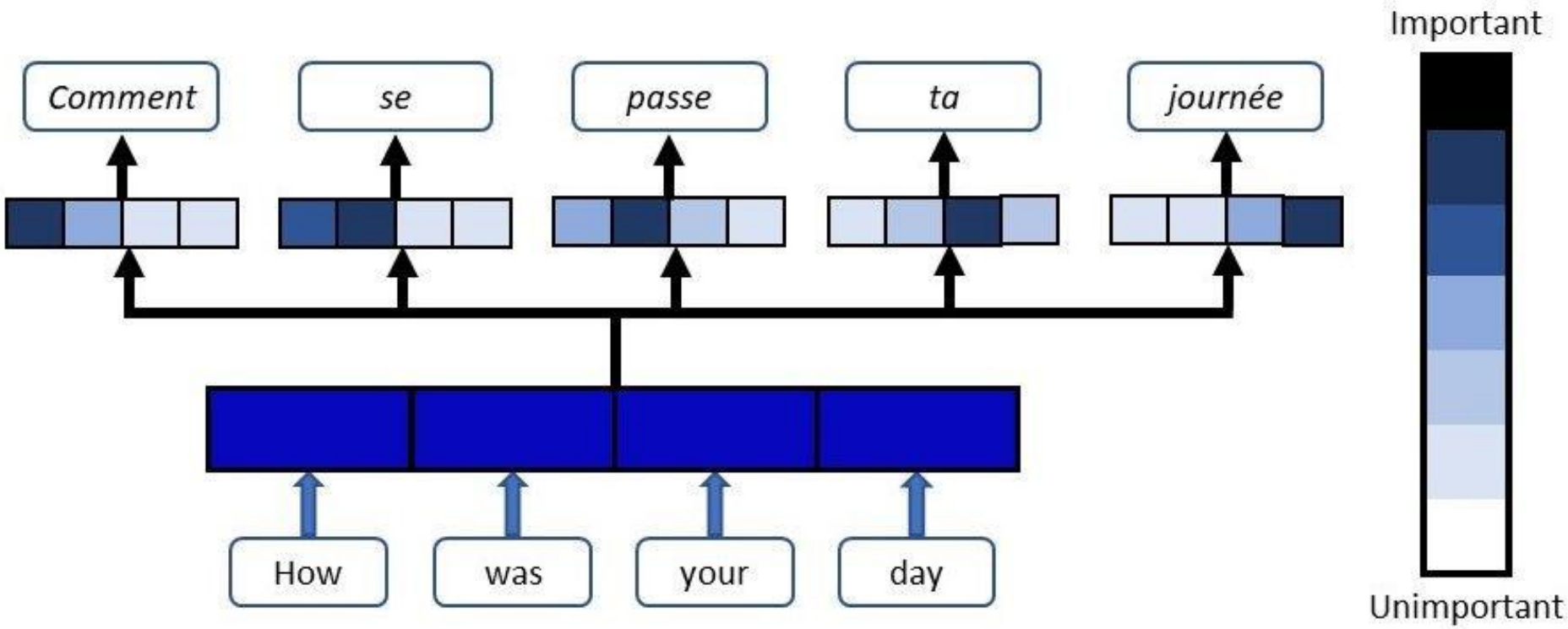
Hidden Layers

Output Layer

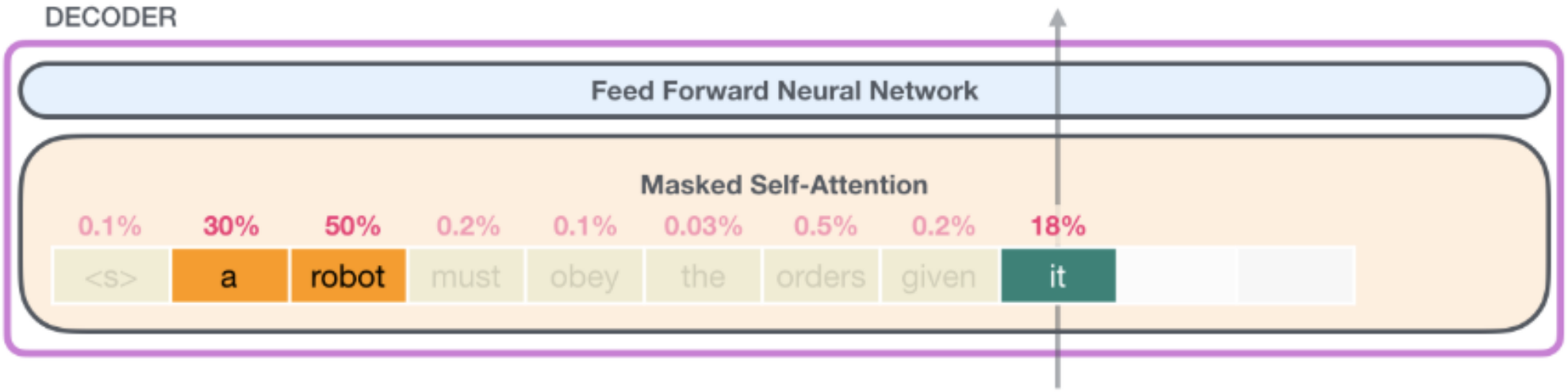
Transformers



Attention


Comment
se
passe
ta
journée
How
was
your
day
Important
Unimportant

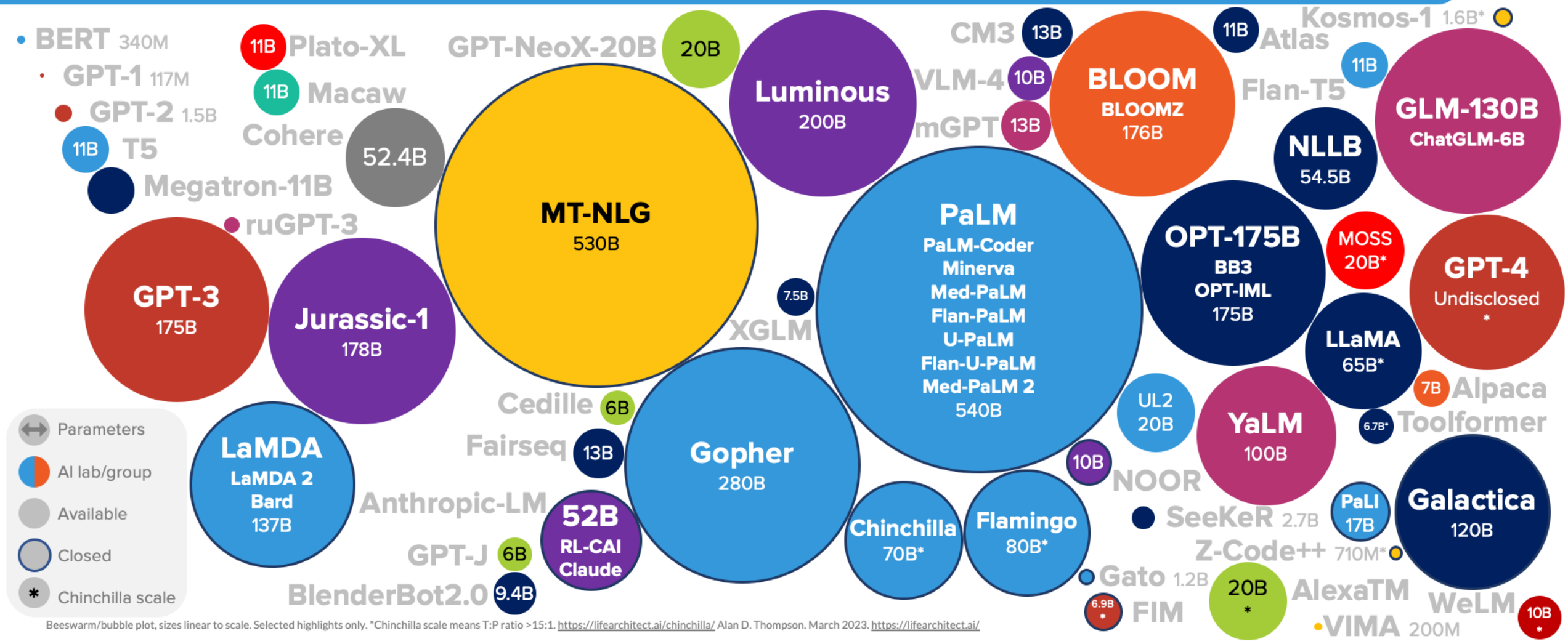
Attention


DECODER
Feed Forward Neural Network
Masked Self-Attention
0.1%
30%
50%
0.2%
0.1%
0.03%
0.5%
0.2%
18%
<s>
a
robot
must
obey
the
orders
given
it

Atentie!

- https://www.menti.com/
- cod: **5193 0229**

LANGUAGE MODEL SIZES TO MAR/2023



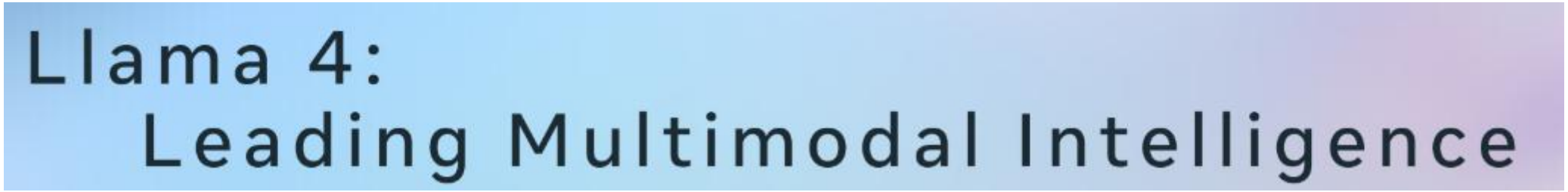
Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

Model	Lab	Playground	Parameters (B)	Tokens trained (B)	Ratio Tokens:Params (Chinchilla scaling≥20:1)	ALScore "ALScore" i Sqr Root of	MMLU	MMLU -Pro	GPQA	HLE	Training dataset	Announced ▼	Public?	Paper / Repo	Arch	Tags
AuroraGPT (ScienceGF	Argonne National L	https://lifearchite	*2000*	*30000*	15:1							*TBA*				
DeepSeek-R2	DeepSeek-AI	https://www.reut										*TBA*			*MoE*	*Reasoning* *SOTA*
ERNIE 5	Baidu	https://lifearchite										*TBA*				
Grok-4	xAI	https://lifearchite										*TBA*			*MoE*	*Reasoning* *SOTA*
GPT-5	OpenAI	https://lifearchite	*5400*	*114000*	22:1							*TBA*			*MoE*	*SOTA*
GPT-6	OpenAI	https://lifearchite										*TBA*				*SOTA*
Llama 4 Reasoning	Meta AI	https://ai.meta.cc										*TBA*		https://ai.m	MoE	SOTA Reasoning
~~*MAI-1*~~	~~Microsoft~~	~~https://arstechni~~	~~500~~	~~*10000*~~	~~20:1~~	~~7.5~~						~~*TBA*~~		~~https://ww~~	~~*Dense*~~	
o4	OpenAI	https://lifearchite										*TBA*				*Reasoning* *SOTA*
Llama 4 Behemoth	Meta AI	https://ai.meta.cc	2000	30000	15:1	25.8		82.2	73.7			Apr/2025		https://ai.m	MoE	SOTA
Llama 4 Maverick	Meta AI	https://ai.meta.cc	400	22000	55:1	9.9		80.5	69.8			Apr/2025		https://ai.m	MoE	SOTA
Qwerky-72B	Featherless AI	https://featherles	72	18000	250:1	3.8	77.46				synthetic, web-scale	Apr/2025		https://featl	Dense	
Agentic-Tx	Google DeepMind	https://github.co	*200*	*20000*	100:1	6.7			62.4	14.5	synthetic, web-scale	Mar/2025		https://stor	*MoE*	
TxGemma	Google DeepMind	https://huggingfa	27	14000	519:1	2.0					synthetic, web-scale	Mar/2025		https://stor	*MoE*	Reasoning
Gemini 2.5 Pro	Google DeepMind	https://aistudio.g	*200*	*20000*	100:1	6.7			84	18.8	synthetic, web-scale	Mar/2025		https://blog	*MoE*	Reasoning SOTA
DeepSeek-V3 0324	DeepSeek-AI	https://chat.deep	685	14800	22:1	10.6		81.2	68.4		synthetic, web-scale	Mar/2025		https://hugg	MoE	SOTA
Llama-3.3-Nemotron-	NVIDIA	https://huggingfa	49	*15040*	307:1	2.9			66.67		web-scale	Mar/2025		https://buil	Dense	Reasoning
EXAONE Deep	LG	https://huggingfa	32	6500	204:1	1.5	83	74	66.1		web-scale	Mar/2025		https://arxi	Dense	Reasoning
Mistral Small 3.1	Mistral	https://huggingfa	24	*8000*	334:1	1.5	81.01	56.03	37.5		web-scale	Mar/2025		https://mist	Dense	
ERNIE 4.5	Baidu	https://yiyan.bai						*79*	*55*		synthetic, web-scale	Mar/2025		https://ww	*MoE*	
X1	Baidu	https://yiyan.bai									synthetic, web-scale	Mar/2025		https://ww	*MoE*	Reasoning
OLMo 2 32B	Allen AI	https://playgrour	32	6400	200:1	1.5	78				synthetic, web-scale	Mar/2025		https://alle	Dense	
Command A	Cohere	https://dashboar	111	*8000*	73:1	3.1	85				synthetic, web-scale	Mar/2025		https://hugg	Dense	
Gemini Robotics	Google DeepMind		*200*	*20000*	100:1	6.7		*79.1*	*64.7*		synthetic, web-scale	Mar/2025		https://stor	*MoE*	

https://lifearchitect.ai/

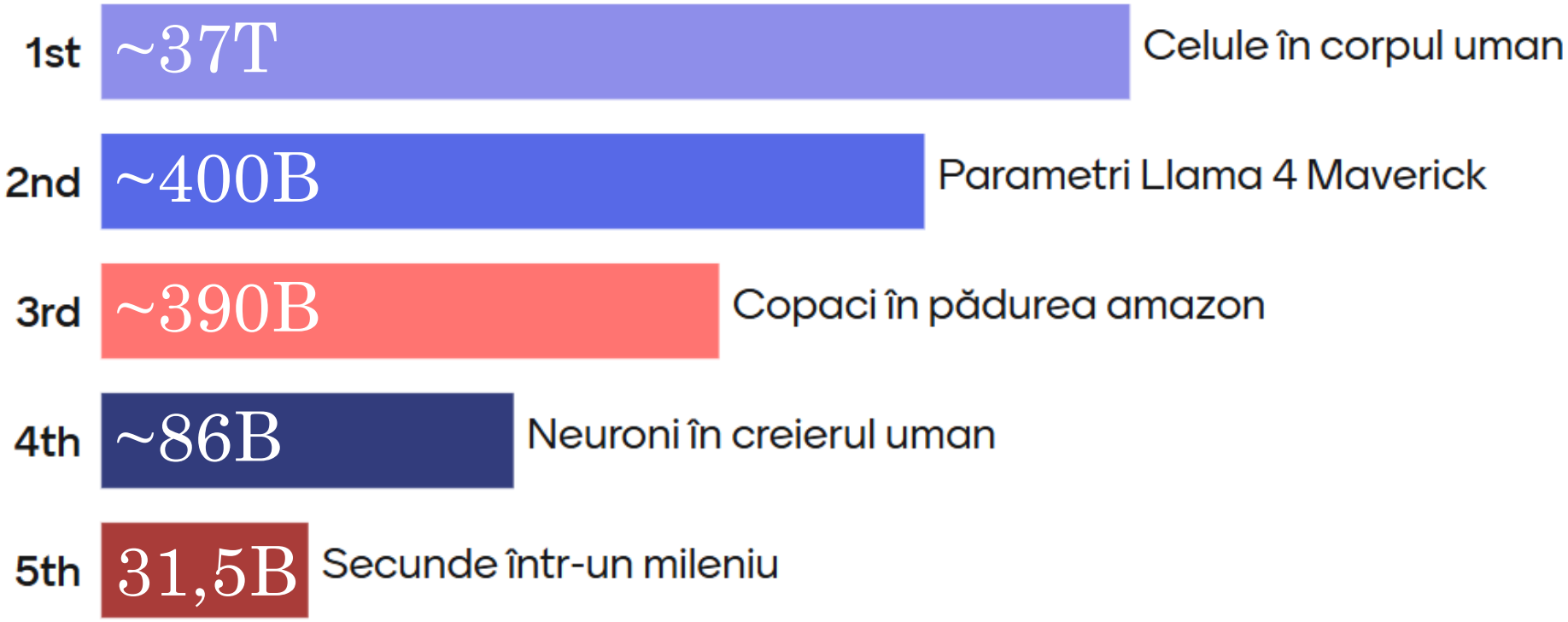
Exemplu LLM

- Llama 4 Maverick
- Dezvoltat de Meta
- 400 **MILIARDE** de parametri
- menti.com cod: **5193 0229**

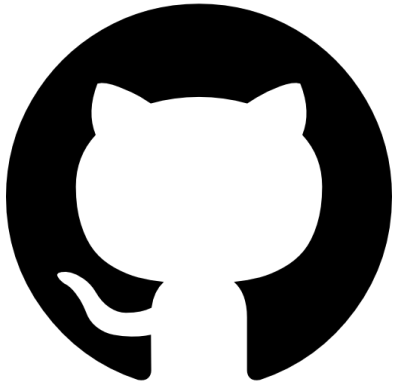


Big numbers

- https://www.menti.com/
- cod: **5193 0229**


1st
~37T
Celule în corpul uman
2nd
~400B
Parametri Llama 4 Maverick
3rd
~390B
Copaci în pădurea amazon
4th
~86B
Neuroni în creierul uman
5th
31,5B
Secunde într-un mileniu

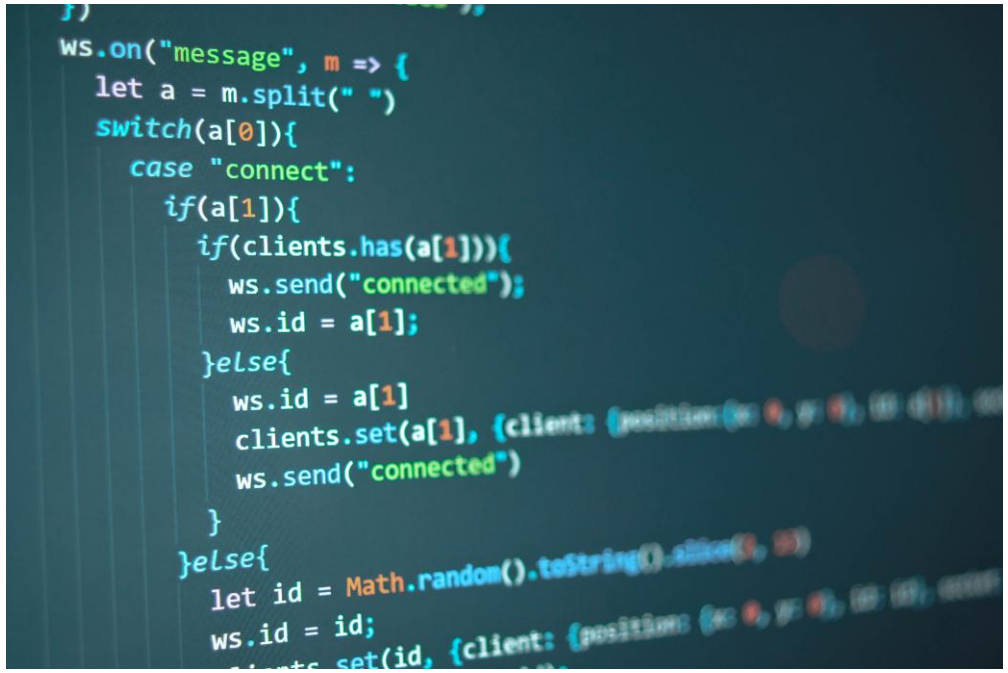
Set de date



Ce mai poate fi generat așa?

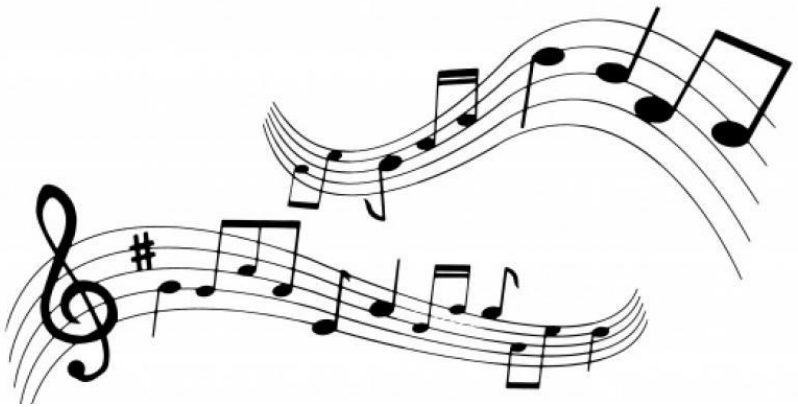
menti.com cod: **5193 0229**

Muzică

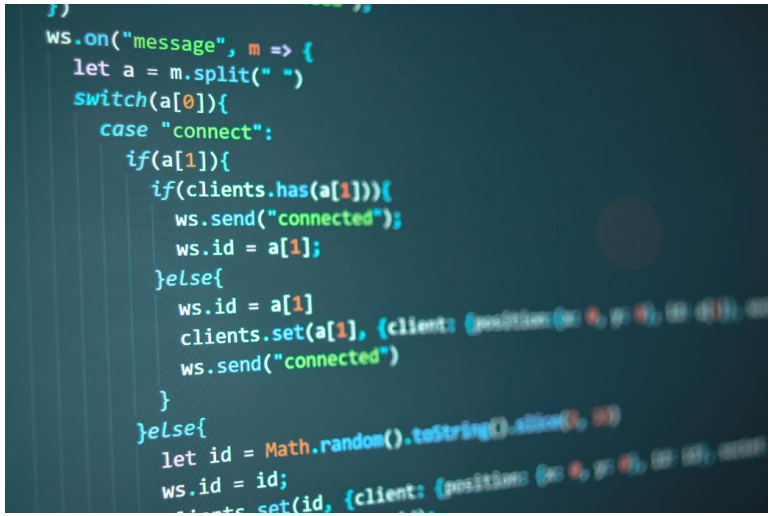


Cod

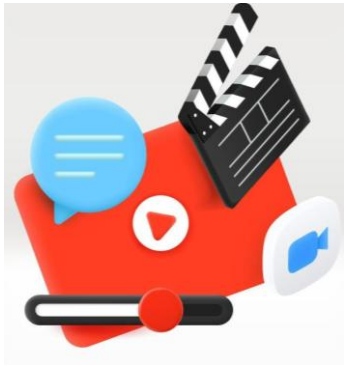
Ce mai poate fi generat așa?



suno.com



Copilot \ Claude \ LLMs



sora.com

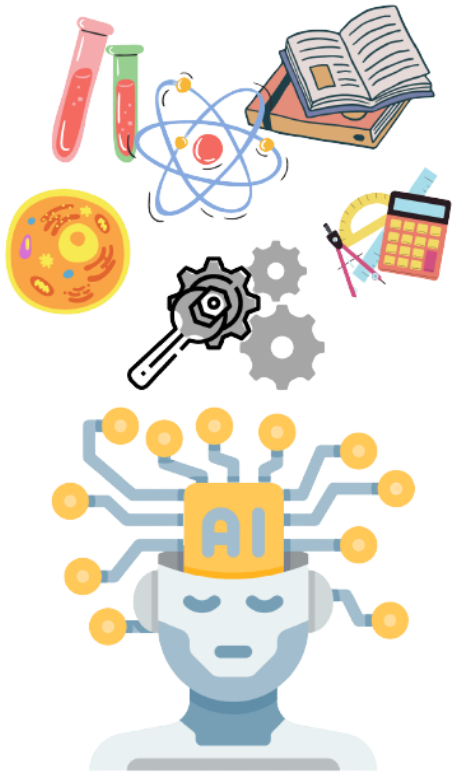
Generare video

- https://www.menti.com/
- cod: **5193 0229**

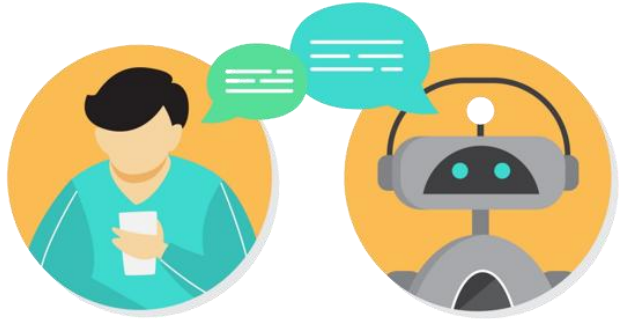
Antrenare LLM

- **Pasul 1:** Antrenare inițială (Pre-training)
  - **Scop:** înțelegerea limbajului natural
  - Prezicerea următorului cuvând din text
- **Pasul 2:** Antrenare de bază (Training alignment)
  - **Scop:** dobândirea de competențe specifice
  - Perechi Prompt X + Răspuns așteptat Y
- **Pasul 3:** Reinforcement Learning
  - **Scop:** îmbunătățirea calității umane
  - Ordonarea răspunsurilor + antrenarea unui model capabil să facă singur ordonarea

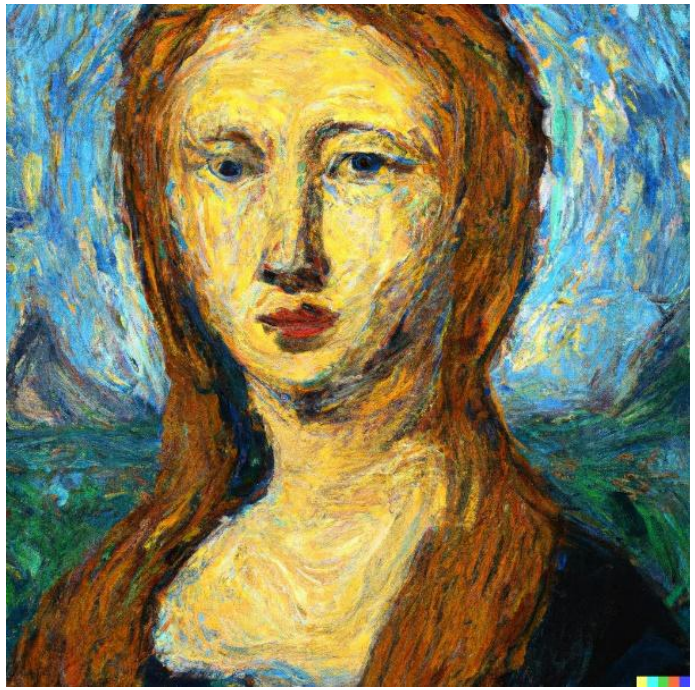
Potențial Generative AI



Pseudo-expert în mai multe domenii



Mod interactiv de a afla informații

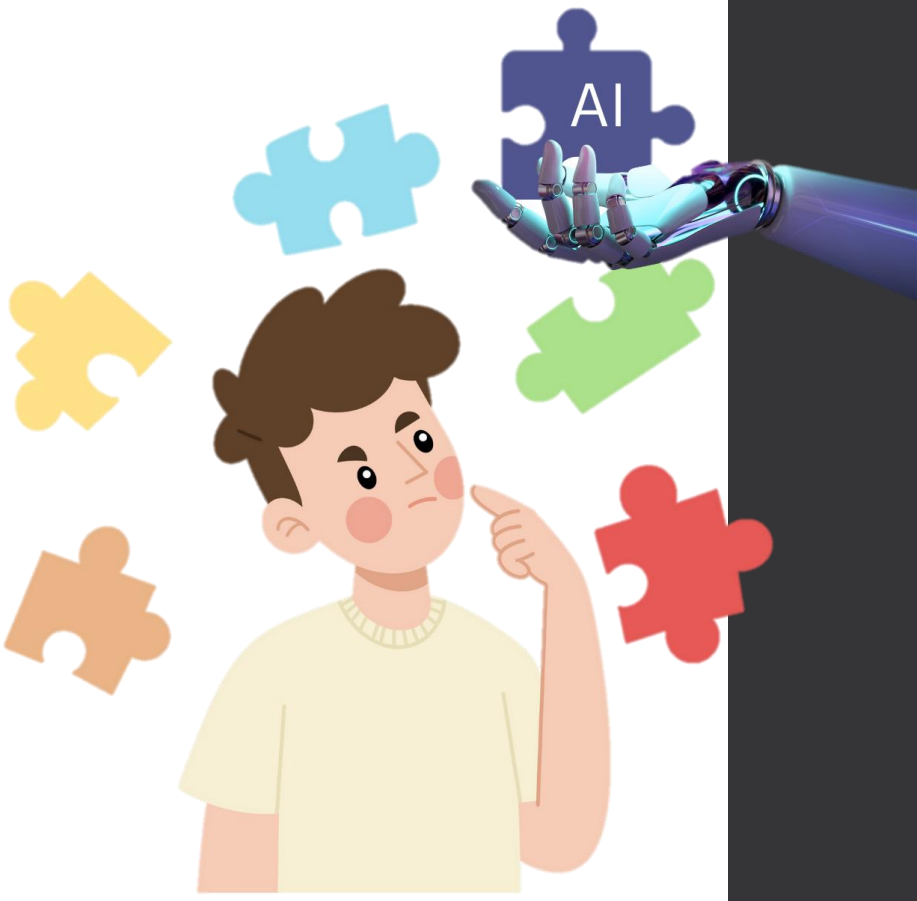


Inspirație creativă

Pericole Generative AI

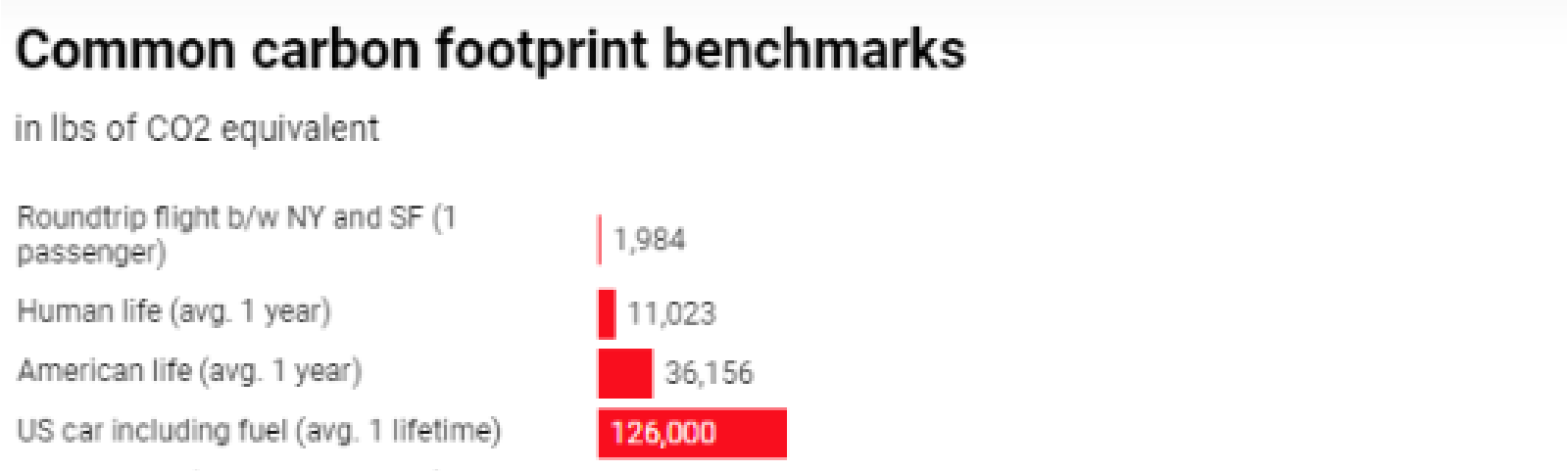
Greșeli, bias, halucinare

Gândire de grup



Are potențialul de a distruge gândirea critică

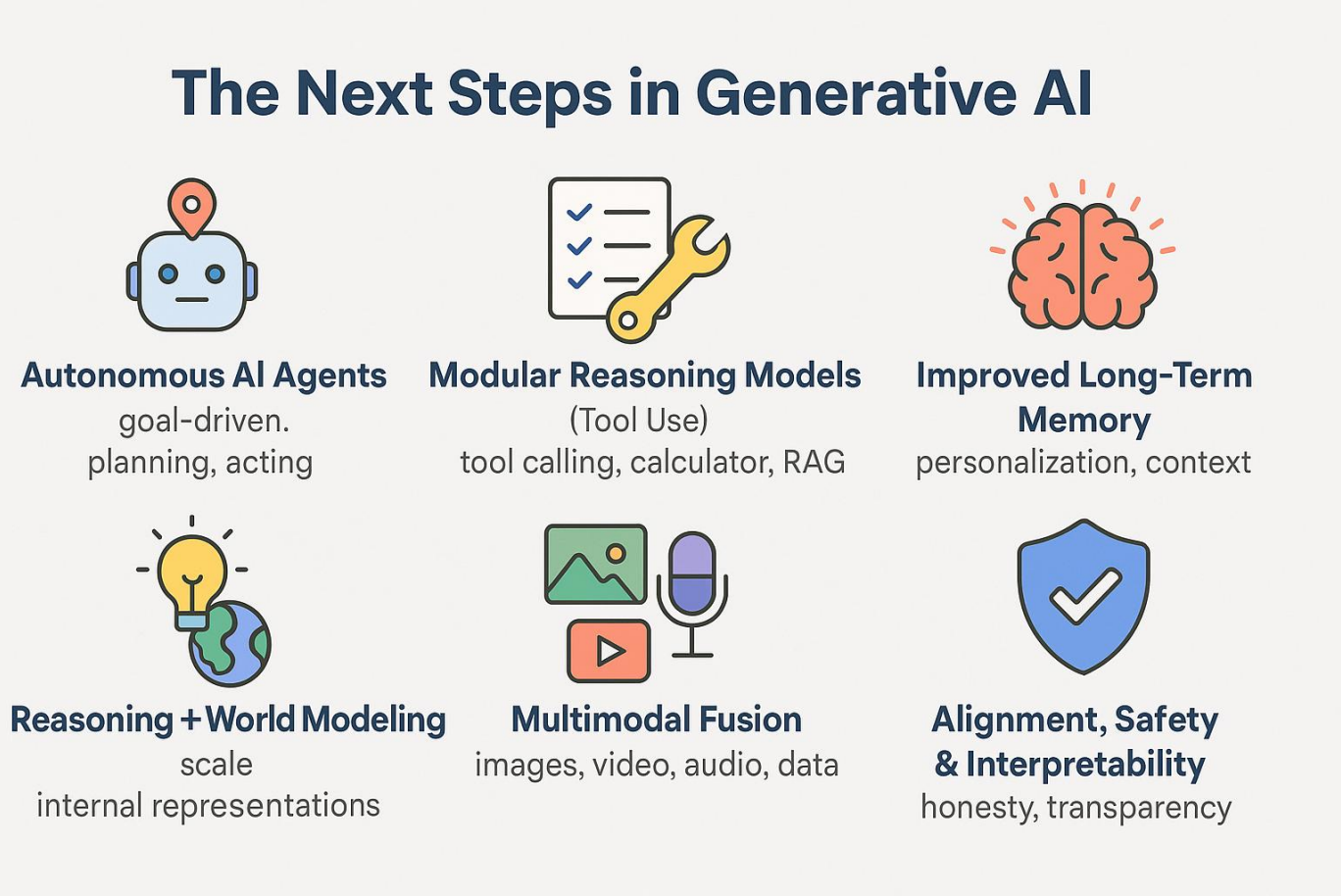
Probleme Generative AI



Ce urmează în Generative AI?

- Continuare\Extindere multi-modal LLMs
- Longer memory
- Autonomous AI Agents
- Reasoning

Ce urmează în Generative AI?



Imagine creată de ChatGPT

Resurse utile

- Karpathy Zero to Hero GPT: https://www.youtube.com/watch?v=kCc8FmEb1nY&list=PLAqhIrjkxbuWI23v9cThsA9GvCAUhRvKZ&index=7
- 3Blue1Brown: https://www.youtube.com/watch?v=wjZofJX0v4M
- Huggin face LLM course: https://huggingface.co/learn/llm-course/chapter1/1
- The annotated Transformer: https://nlp.seas.harvard.edu/annotated-transformer/
- Illustrated GPT2: https://jalammar.github.io/illustrated-gpt2/
- Running a LLM: https://huggingface.co/docs/transformers/llm_tutorial
- Visual Question Answering: https://huggingface.co/docs/transformers/main/en/tasks/visual_question_answering

Sfărșit partea I

Mulțumesc!