

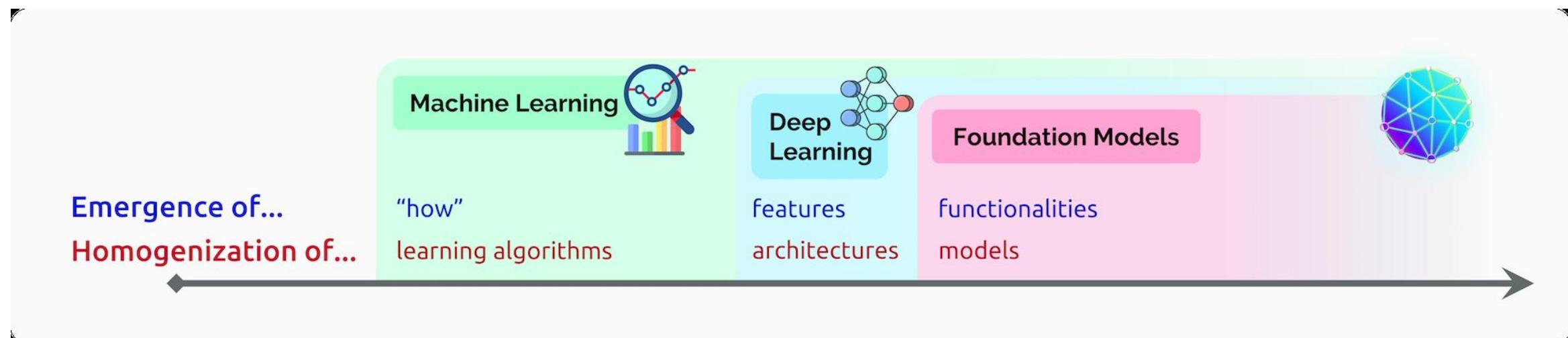
GENERATIVE AI & FOUNDATION MODELS

FOUNDATION MODELS

- A foundation model is any model that is trained on broad data (generally using **self-supervision at scale**) that can be **adapted** (e.g., **fine-tuned**) to a wide range of downstream tasks; current examples include BERT, GPT-3, and CLIP.
- Foundation models are not new from a technological point of view:
 - Based on deep neural networks
 - Self-supervised learning, both of which have existed for decades
- The sheer scale and scope of foundation from the last few years have stretched our imagination of what is possible
 - GPT-3 has 175 billion parameters
 - Can be adapted via **natural language prompts** to do a passable job on a wide range of tasks despite **not being trained explicitly** to do many of those tasks

EMERGENCE AND HOMOGENIZATION

- Emergence means that the **behavior** of a system is **implicitly induced** rather than **explicitly constructed**.
- Emergence can be both the **source of scientific excitement and anxiety** about unanticipated consequences.
- Homogenization indicates the **consolidation of methodologies** for building machine learning systems across a wide range of applications
- Homogenization provides strong leverage towards many tasks but also creates single points of failure.

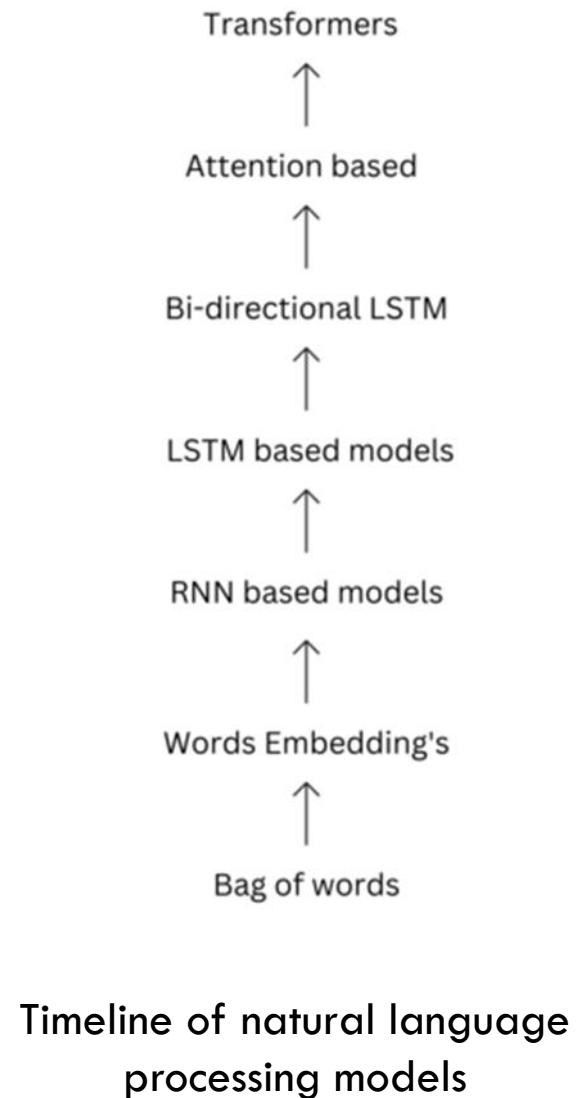


FOUNDATION MODELS

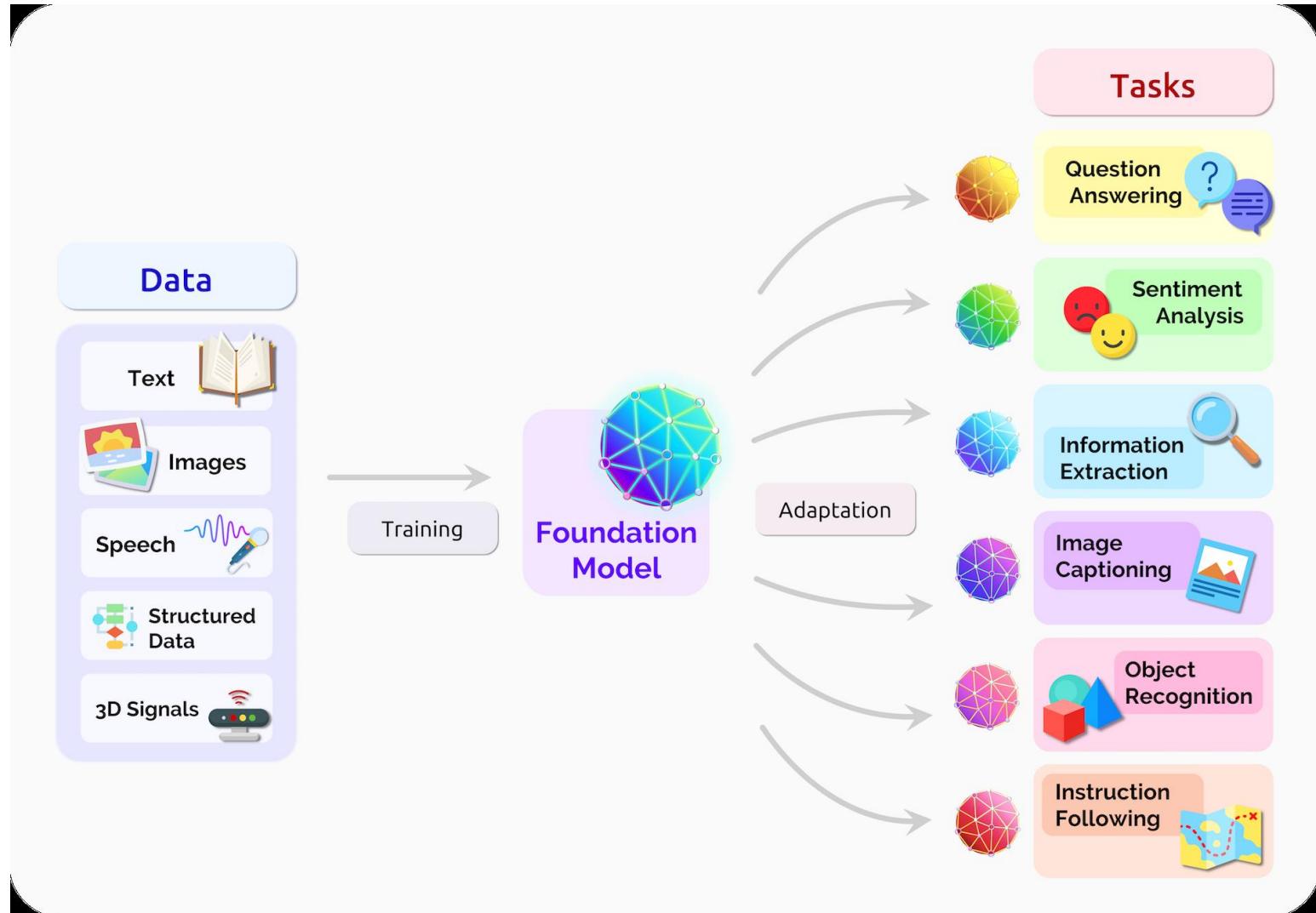
- Foundation models are enabled by **transfer learning** and **scale**.
- **Transfer Learning:** idea is to take the “knowledge” learned from one task (e.g., object recognition in images) and apply it to another task (e.g., activity recognition in videos).
- Within deep learning, **pre-training** is the dominant approach to transfer learning:
 - a model is trained on a **surrogate** task (often just as a means to an end) and then **adapted** to the **downstream** task of interest via fine-tuning.
- Scale makes foundation models powerful:
 - The Three ingredients of Scale:
 - Improvements in computer hardware — e.g., GPU throughput and memory have increased 10× over the last four years (2017-2021)
 - The development of the **Transformer model architecture** [Vaswani et al. 2017] that leverages the parallelism of the hardware to train much more expressive models than before
 - The availability of much more training data.

FOUNDATION MODELS

- In self-supervised learning the pre-training task is derived automatically from unannotated data.
- For example, the masked language modeling task used to train BERT [Devlin et al. 2019] is to predict a missing word in a sentence given its surrounding context (e.g., I like ___ sprouts).
- Self-supervised tasks are:
 - More scalable as they only depend on unlabeled data
 - Designed to **force the model** to predict parts of the inputs
 - Makes them richer and potentially more useful than models trained on a more limited label space.



A FOUNDATION MODEL CAN CENTRALIZE THE INFORMATION FROM ALL THE DATA FROM VARIOUS MODALITIES. THIS ONE MODEL CAN THEN BE ADAPTED TO A WIDE RANGE OF DOWNSTREAM TASKS.



TRANSFORMER ARCHITECTURE

All transformers have the same primary components:

Tokenizers, which convert text into tokens.

A single embedding layer, which converts tokens and positions of the tokens into vector representations.

Transformer layers: carry out repeated transformations on the vector representations, extracting more and more linguistic information.

These consist of alternating **attention** and **feedforward** layers.

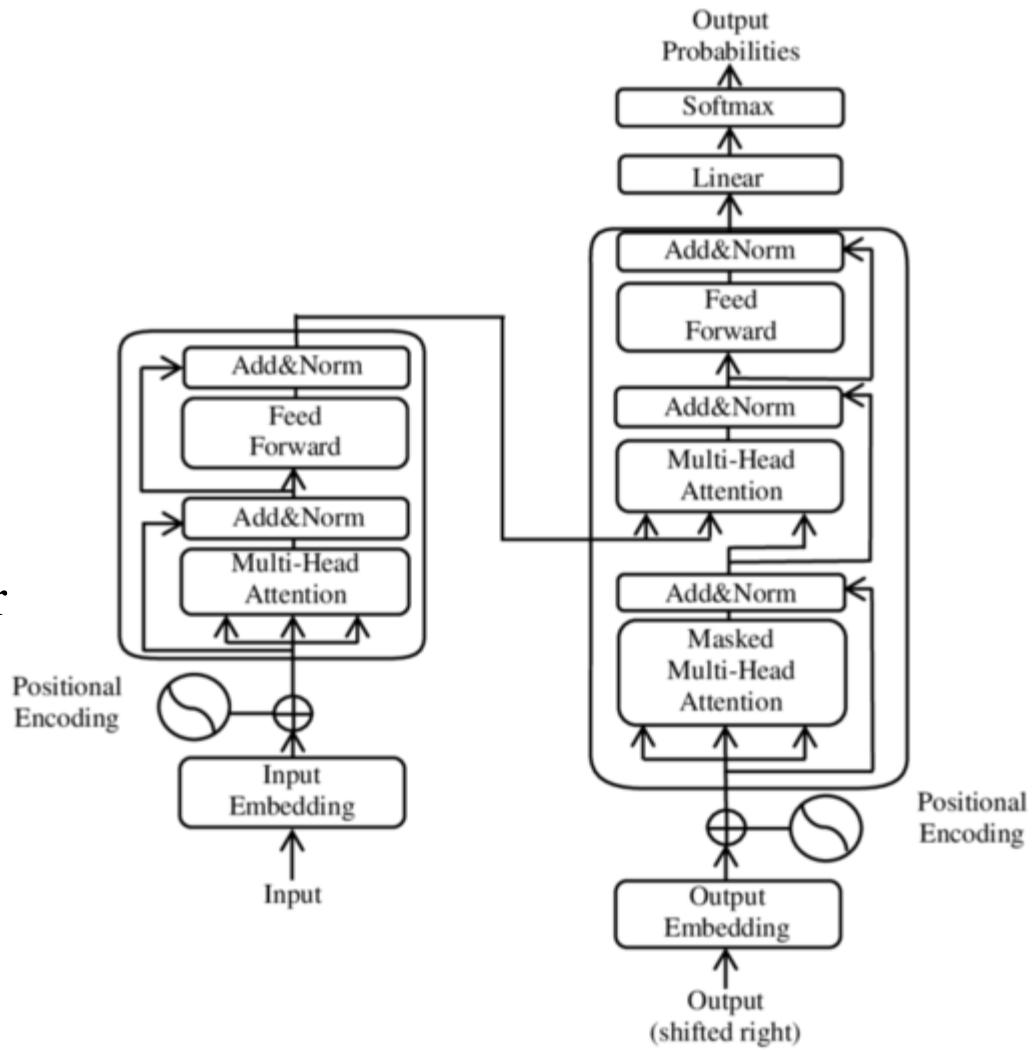
Un-embedding (Optional) layer, which converts the final vector representations back to a probability distribution over the tokens.

Transformer layers can be one of two types:

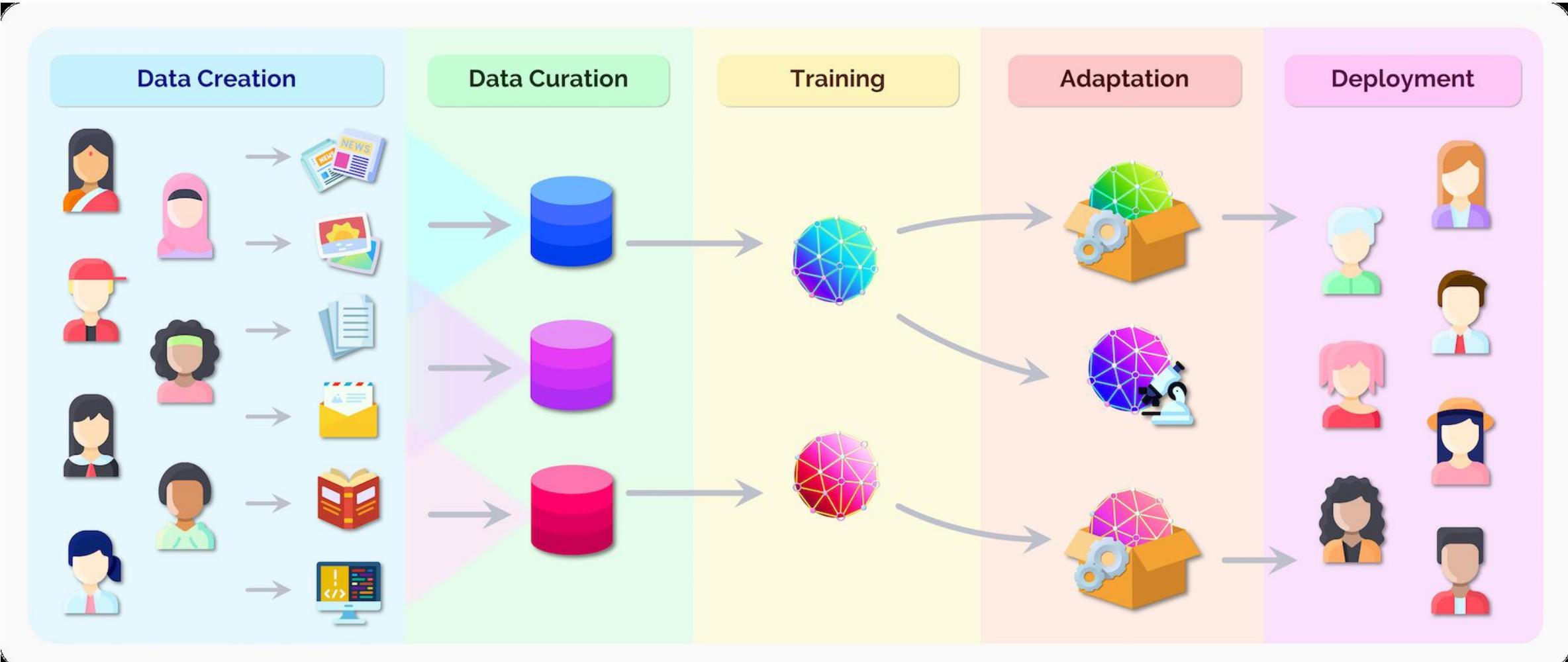
Encoder and decoder: original paper used both encoder and Decoder

Later models included only one type of them.

BERT is an example of an encoder-only model; GPT are decoder-only models.



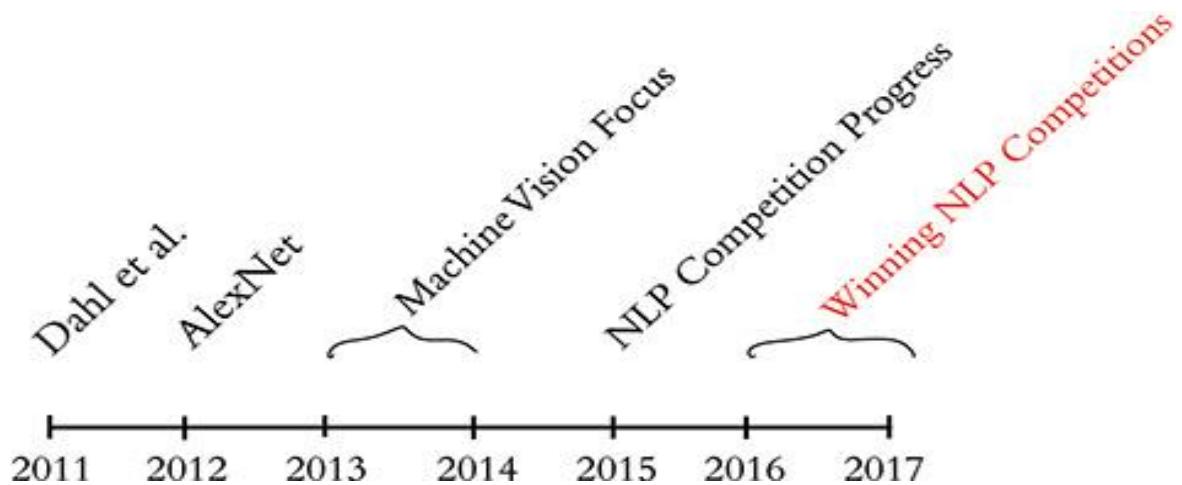
FOUNDATION MODELS ARE PART OF A BROADER ECOSYSTEM THAT STRETCHES FROM DATA CREATION TO DEPLOYMENT



On Large Language Models for Understanding Human Language

Christopher Manning

- <https://www.youtube.com/watch?v=YfXc4OBMnM>



2011: Dahl et al first applied a deep learning algorithm to a large dataset – recognize a substantial vocabulary of words from audio recordings of human speeches

2012-2015: more focus on Machine Vision

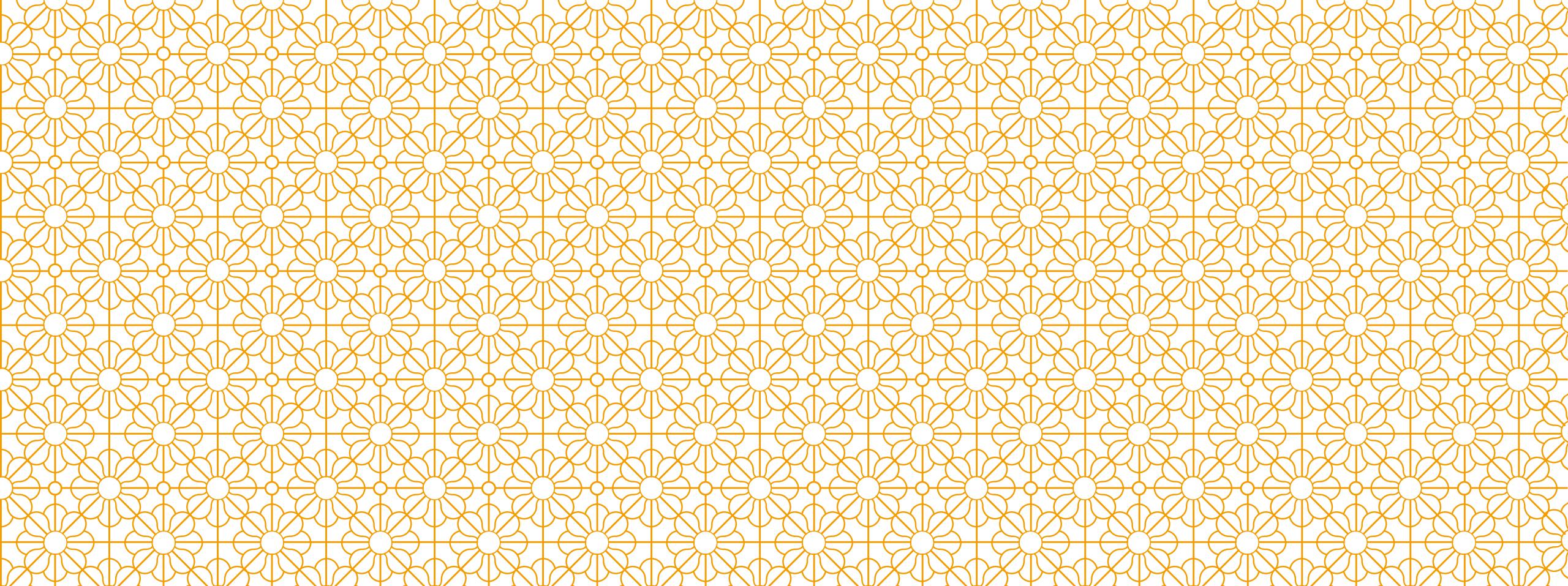
Deep learning based models approached the precision and accuracy of traditional machine learning models with

- Less development time
- Lower computational complexity

Microsoft was able to integrate real-time machine translation software onto mobile phone processors

BRIEF HISTORY

- N-gram models – before neural networks
 - Automatic Speech Recognition, Machine Translation
 - Improved with more tokens:
 - an n-gram model for English with a total of **1.8T n-grams** and noted steady improvements in **BLEU** score on the test set of 1797 Arabic translations as the training data was increased from 13M tokens.
- Pretrained representations of the distribution of words (also called word embeddings)
- Examples: word2vec, GloVe
 - LSTM models such as context2vec and ELMo
 - Applications supported included: question answering, textual entailment, semantic role labeling (SRL), coreference resolution, named entity recognition (NER), and sentiment analysis
 - English and later in other languages
- Increasing the number of model parameters didn't yield noticeable improvements in performances for LSTMs



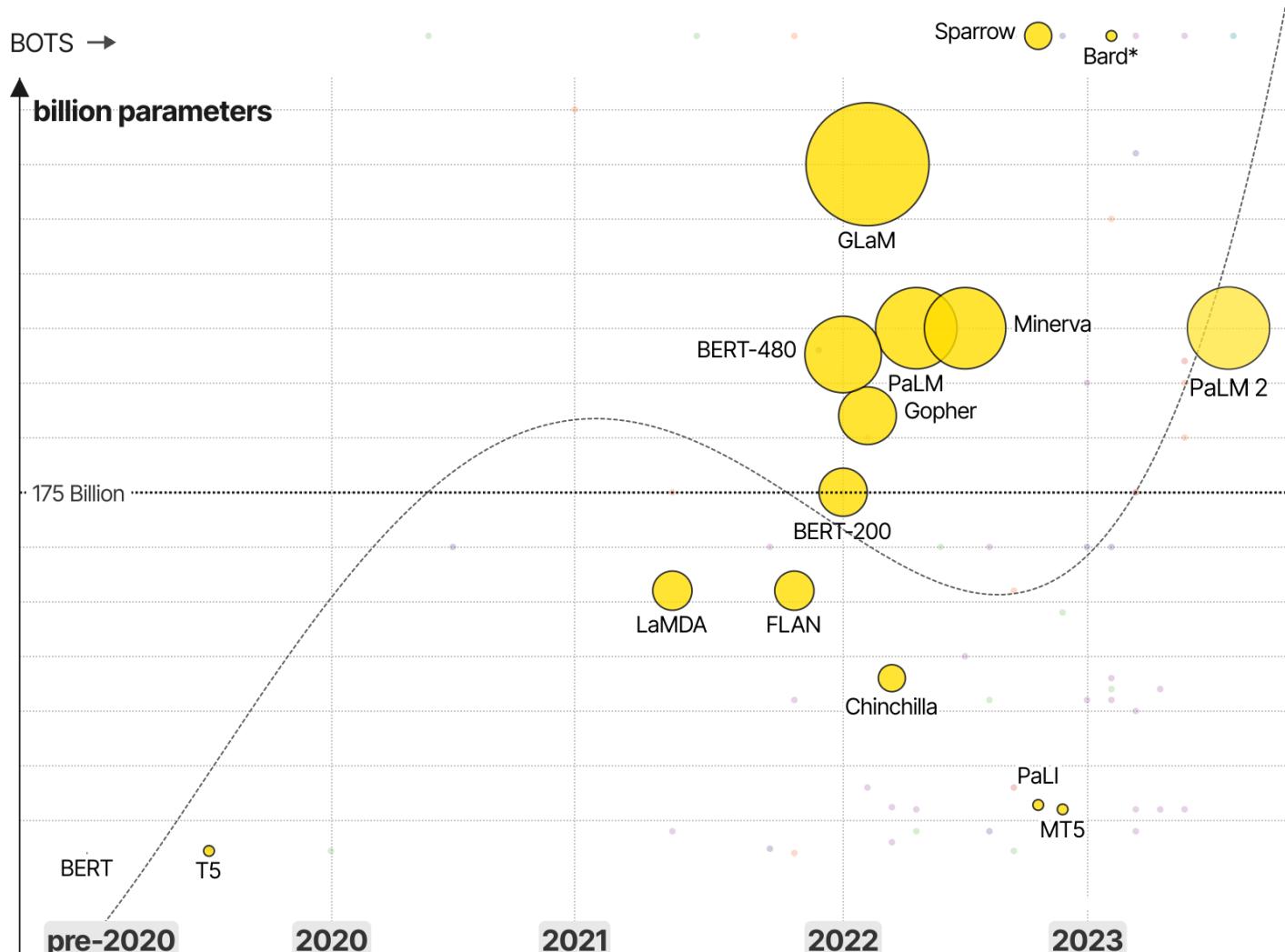
BRIEF HISTORY OF LLMS

BRIEF HISTORY

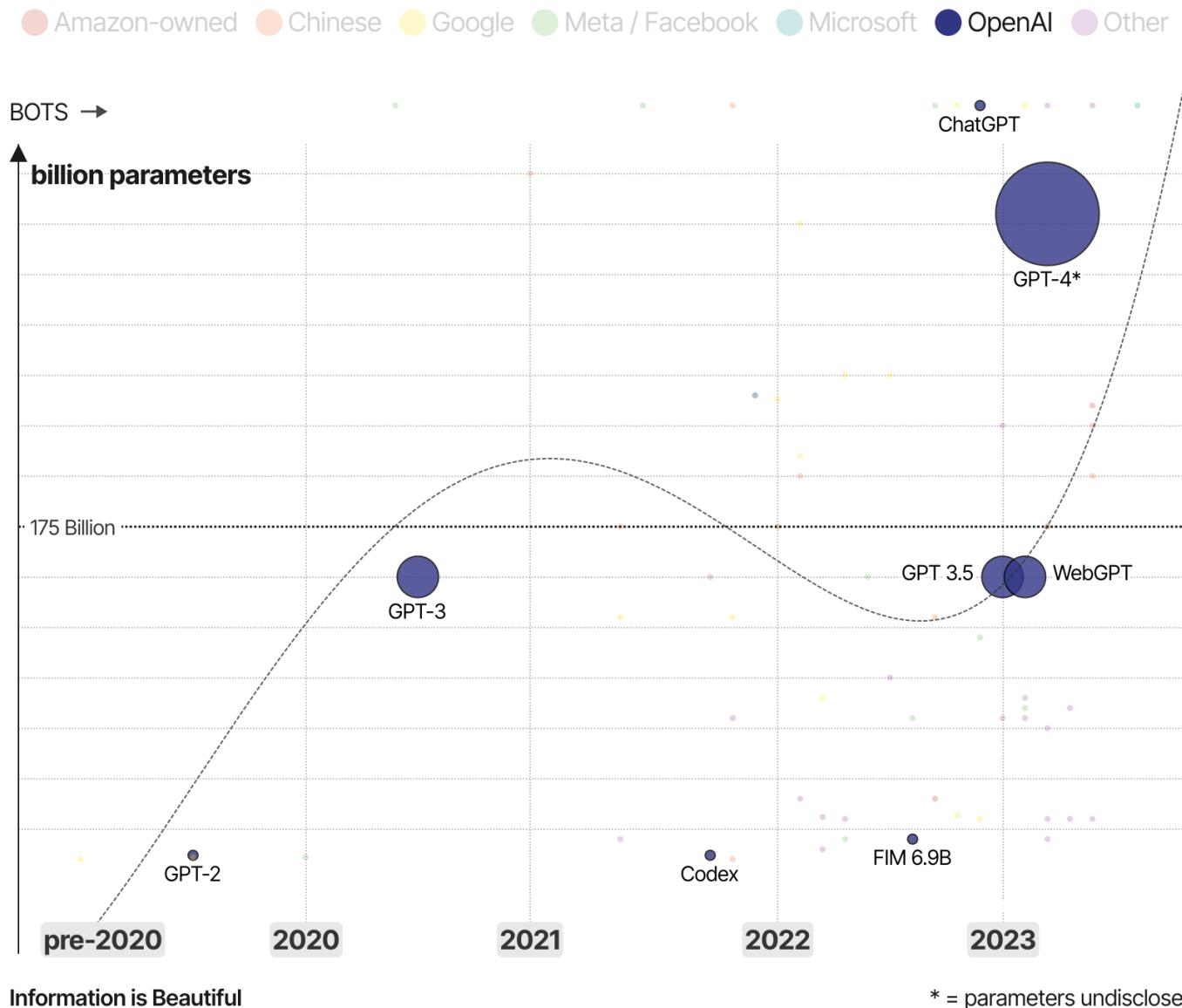
- A **transformer** is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data.
- Transformers were introduced in 2017 by a team at Google Brain
- Now, the model of choice for NLP problems
- Replaced RNN models such as long short-term memory (LSTM).
- Training parallelization allows training on larger datasets.
- Development of pretrained systems such as:
 - BERT (Bidirectional Encoder Representations from Transformers)
 - GPT (Generative Pre-trained Transformer)
- **Training data:** large language datasets, such as the Wikipedia Corpus, Common Crawl, Colossal Clean Crawled Corpus, Pile, etc.
- Transformer models are continuously benefitting from larger architectures and larger quantities of data.
- Transformer models can be fine-tuned for specific tasks

GOOGLE DROVE A BURST OF INNOVATION IN THE LLM SPACE

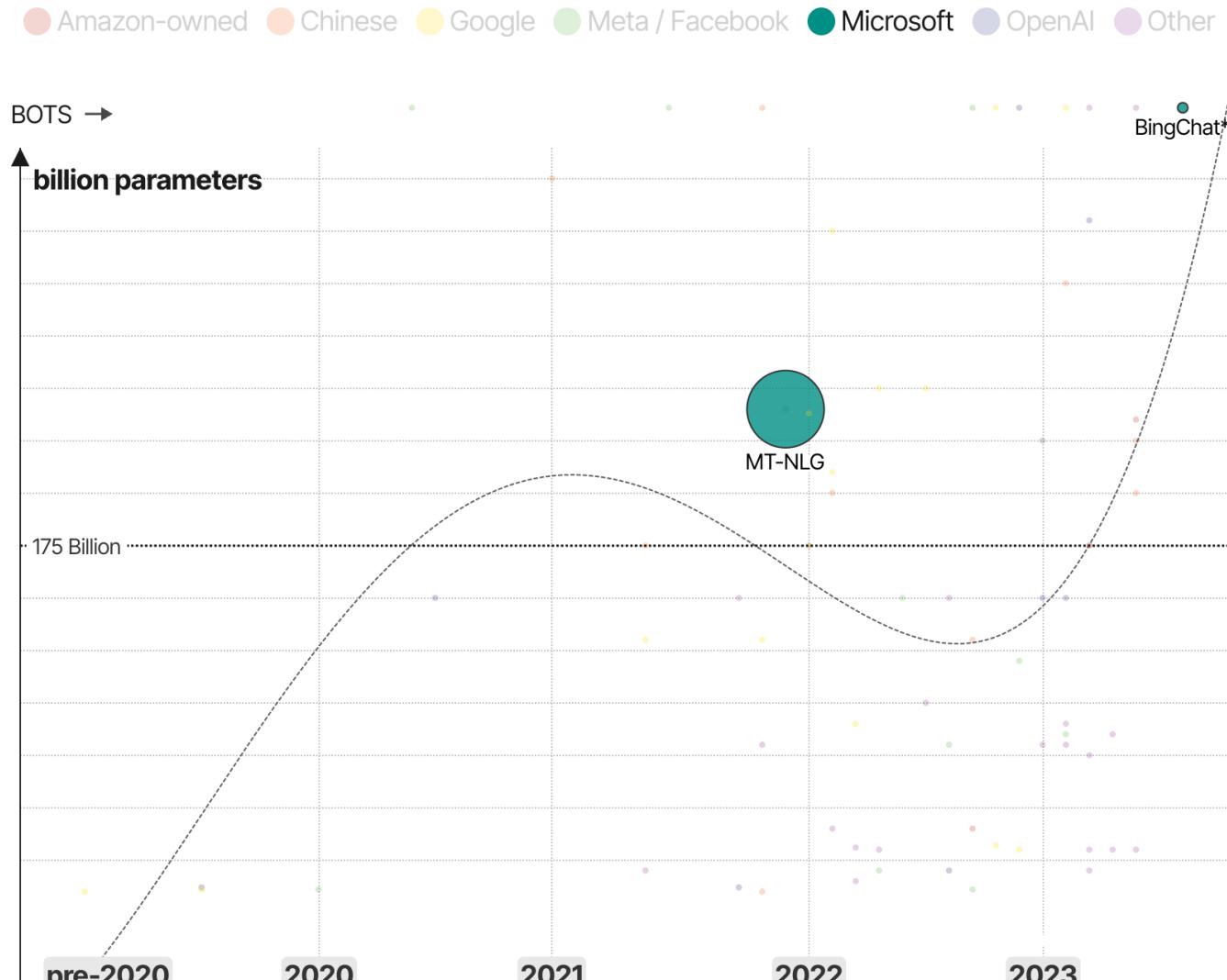
● Amazon-owned ● Chinese ● Google ● Meta / Facebook ● Microsoft ● OpenAI ● Other



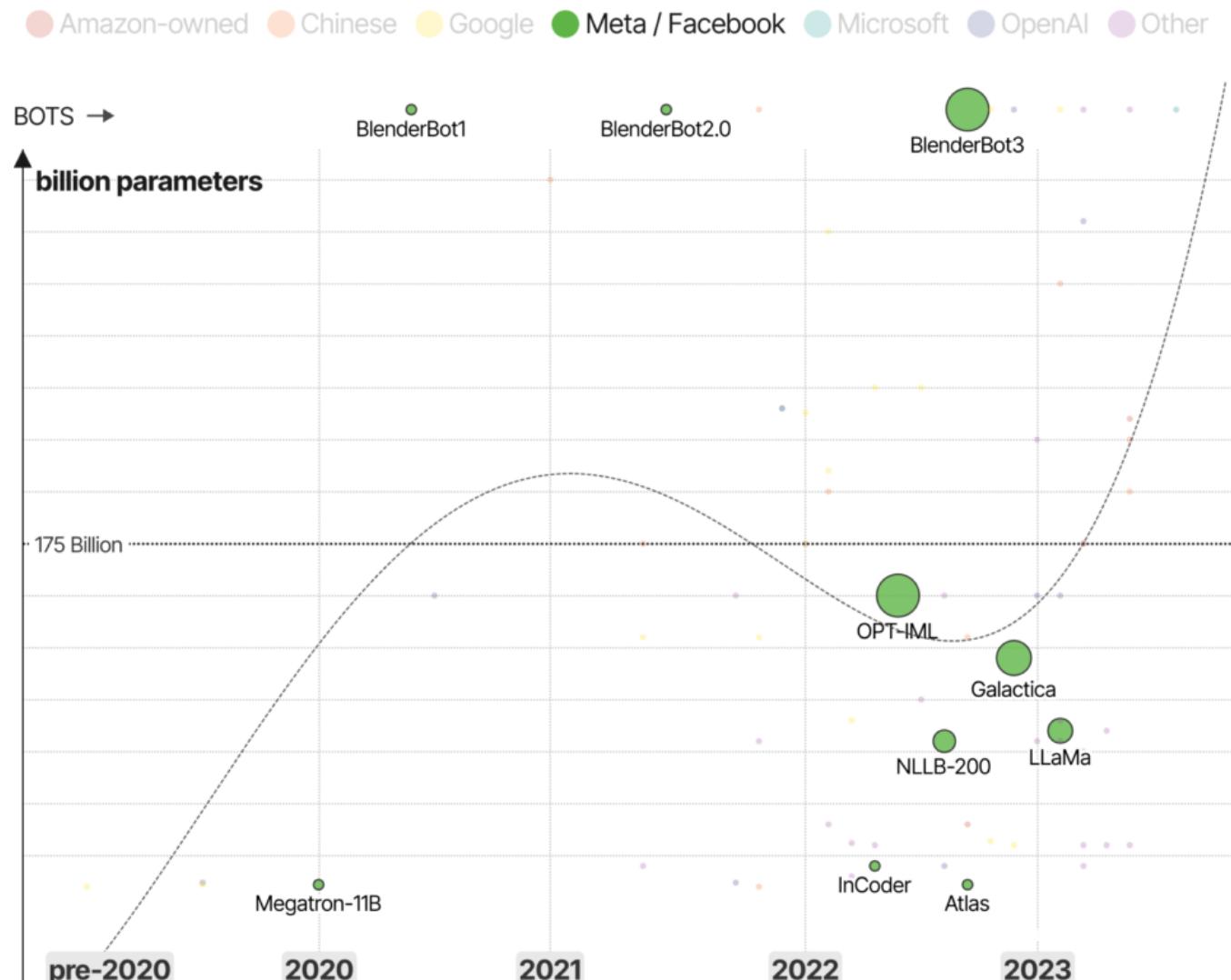
However, OpenAI, creators of ChatGPT, stole the LLM show
They made steady, solid progress over the three years, driving the curve



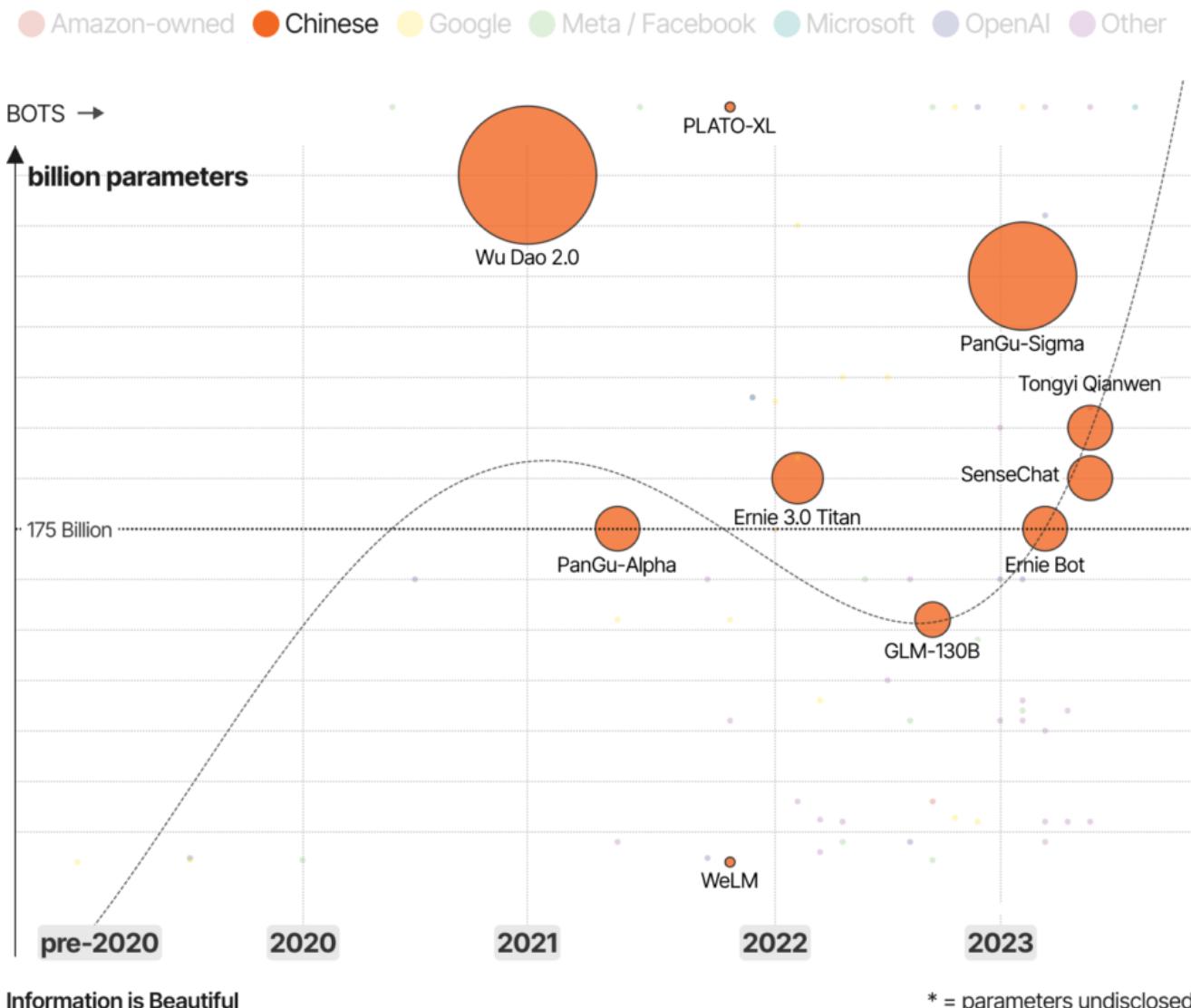
Microsoft weren't directly active in the LLM space with their own research. Instead they invested early and hard in OpenAI (\$1bn in both 2019 and 2021) and that paid serious dividends.



Meta / Facebook also drove significant early innovation in the field...

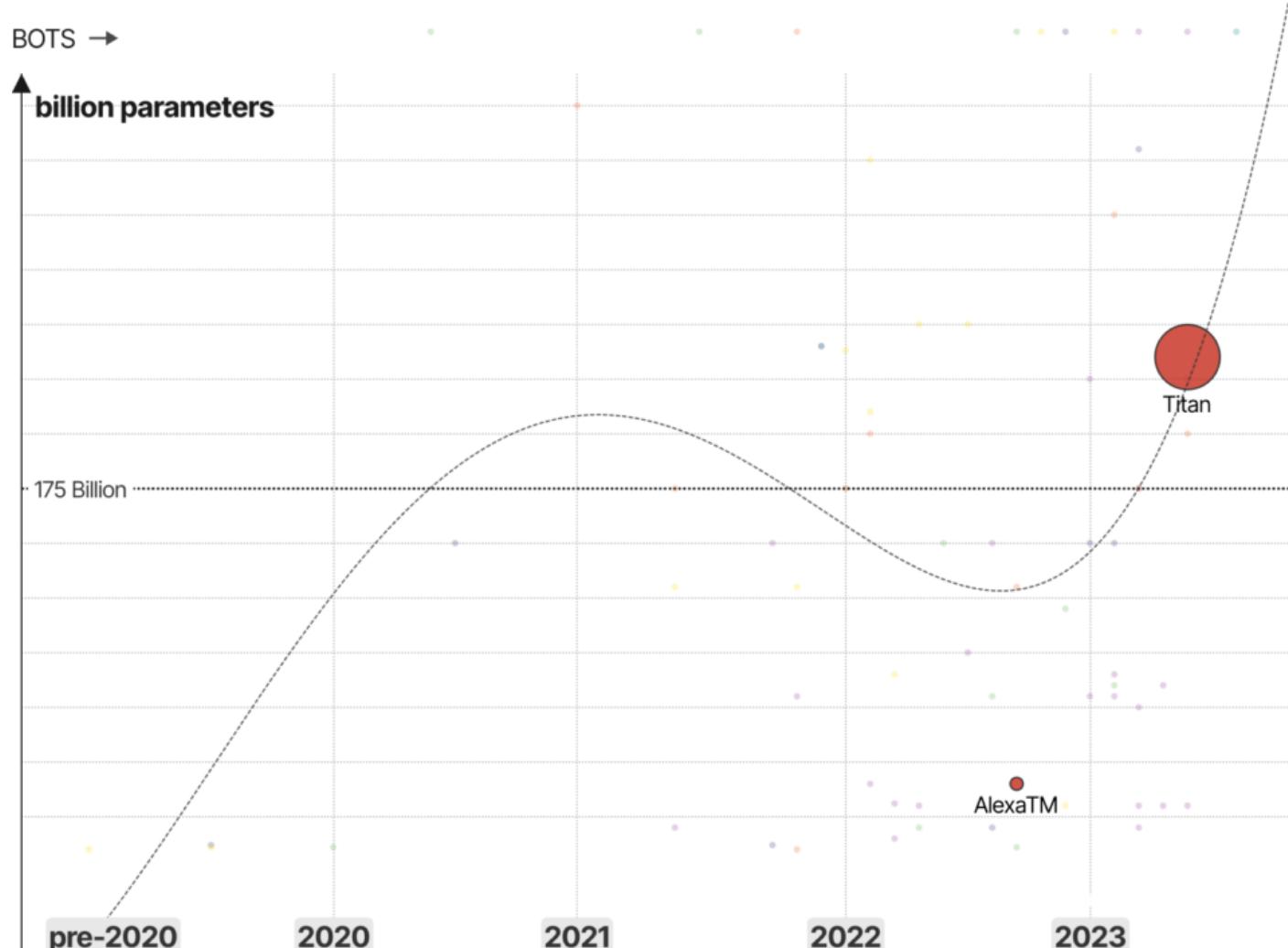


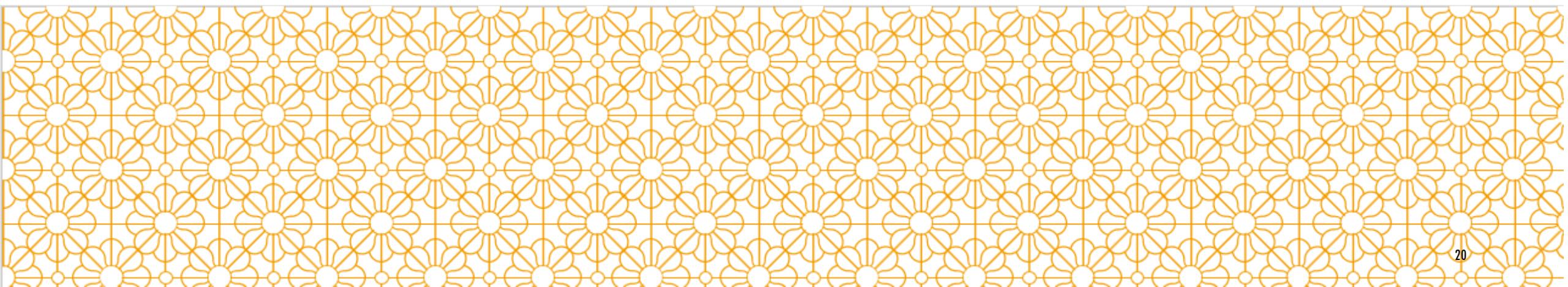
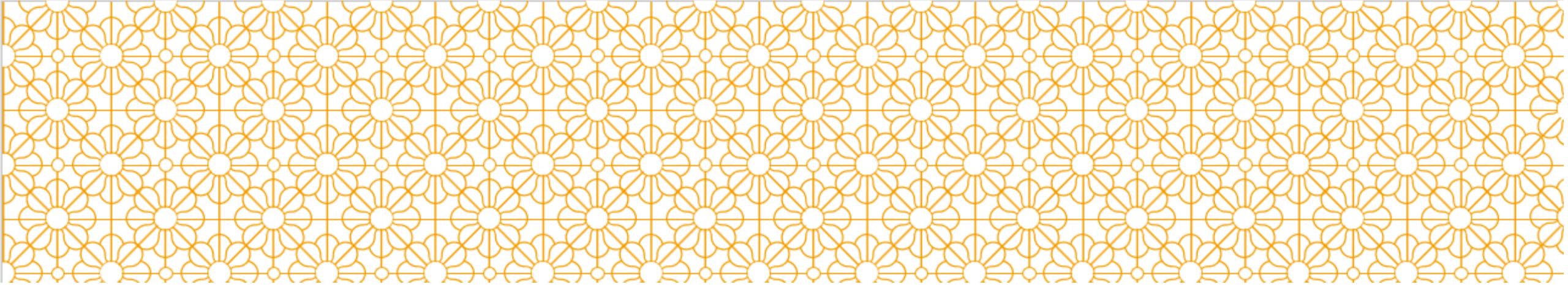
Meanwhile, in the background, China made steady progress. After the advent of ChatGPT, release of Chinese-language LLM's and chatbots have significantly accelerated.



Amazon have steamed in at the end. Too late to the party? Time will tell... Though they have recently invested heavily in Anthropic, creators of impressive LLM Claude

● Amazon-owned ● Chinese ● Google ● Meta / Facebook ● Microsoft ● OpenAI ● Other



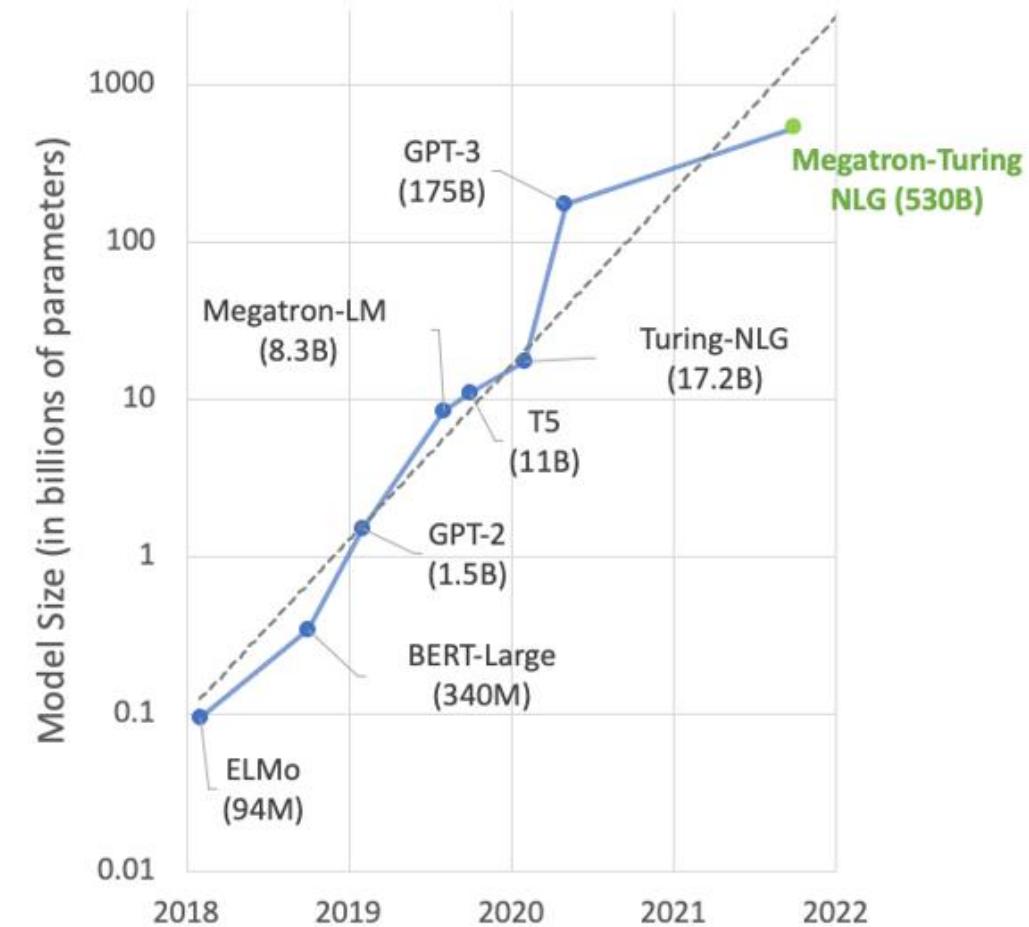


**LANGUAGE MODELS HAVE BECOME BIGGER AND
BIGGER**

LANGUAGE MODELS HAVE BECOME BIGGER AND BIGGER [1]

Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-GEN (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	—
2021	Switch-C [43]	1.57E+12	745GB

Table 1: Overview of recent large language models



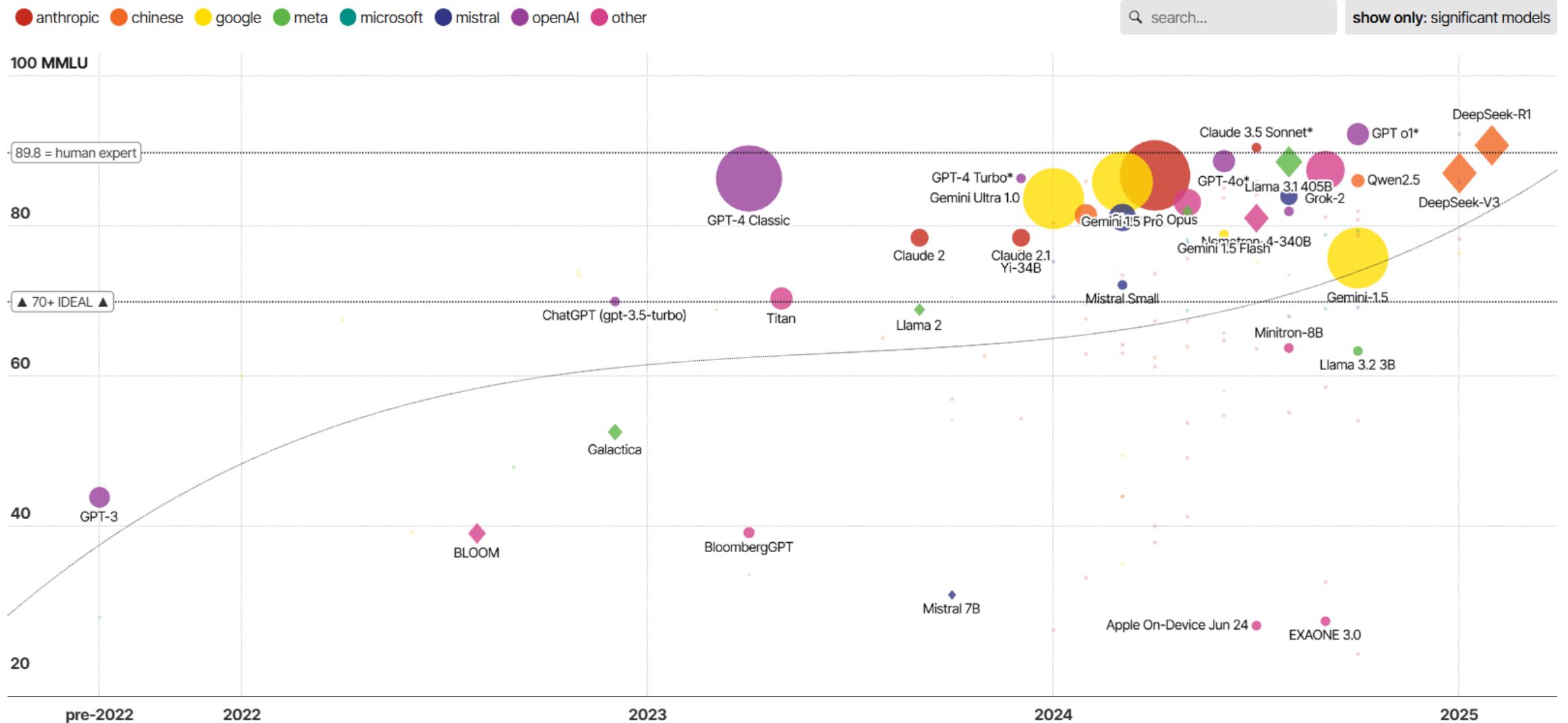
ChatGPT → # of Parameters: 175 Billion

GPT 4 → # of Parameters: 1 Trillion parameters?

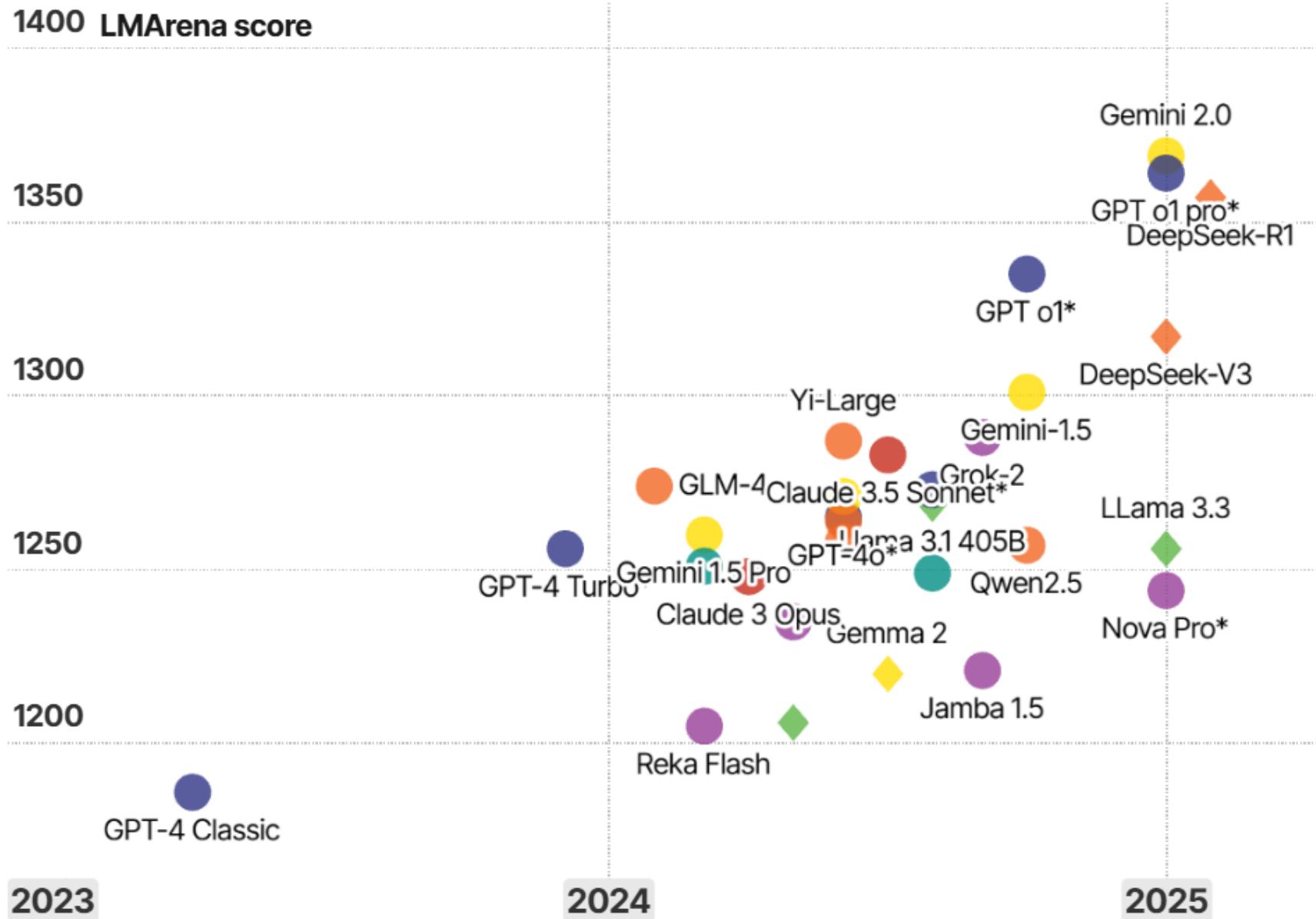
LANGUAGE MODELS HAVE BECOME BIGGER AND BIGGER [2]

LARGE LANGUAGE MODEL	Parameters
Phi-1.5	1.3B
Phi-2	2.7B
Llama2	7B, 13B, or 70B
BloombergGPT	50B
Claude2	130B
GPT-3	175B
GPT-4 “32k”	1.76T

Major Large Language Models (LLMs) ranked by capabilities, sized by billion parameters used for training



Major LLMs ranked by LM Arena Score* LM Arena: AI Benchmarking community with over 1m user votes



LLMS: ISSUES AND CONCERNS

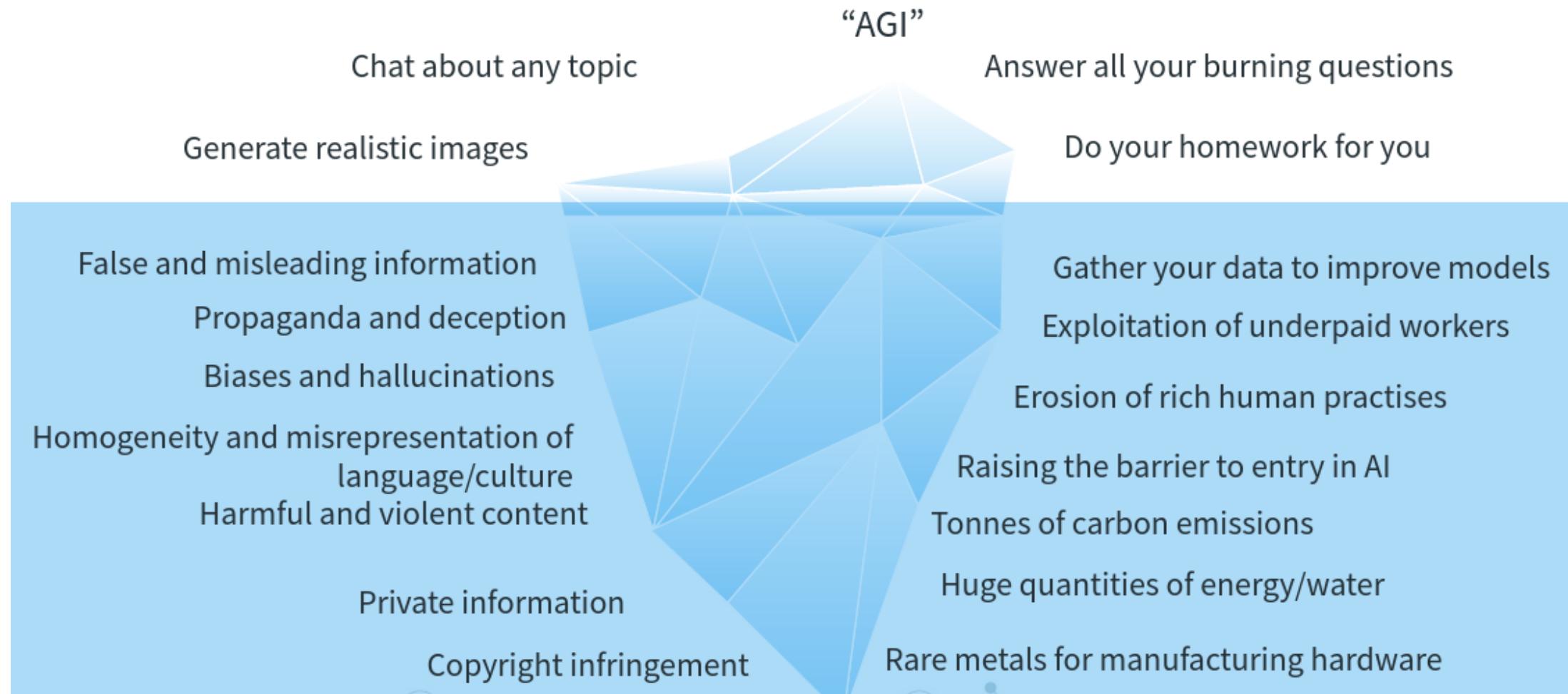
TRENDS AND ISSUES

- ❖ LMs models do not address inclusion problems
- ❖ Over 90% of the world's languages used by more than a billion people currently have little to no support in terms of language technology.
- ❖ Trend in reducing the size of these models using various techniques
 - ❖ Knowledge distillation, quantization, factorized embedding parameterization, and cross-layer parameter sharing, and progressive module replacing
- ❖ Some features of models derived from BERT:
 - ❖ These models maintain and sometimes exceed the performance of the original BERT model,
 - ❖ They are much smaller size
 - ❖ They ultimately still rely on large quantities of data, significant processing and storage capabilities to both hold and reduce the model.

EXPANSION AND CHANGES IN TYPES OF TASKS

- ❖ Moving from n-gram LMs → word vectors distilled from neural LMs → pretrained Transformer LMs
- ❖ Similar expansion and changes in the types of tasks LMs are useful for:
 - ❖ n-gram LMs → **selecting among the outputs** of e.g. acoustical or translation models
 - ❖ LSTM-derived word vectors → **more effective representations of words** (in place of bag of words features) in a variety of NLP tasks involving labeling and classification
 - ❖ Pretrained Transformer models can be retrained on very small datasets (few-shot, one-shot or even zero-shot learning) to perform apparently **meaning-manipulating tasks such as summarization, question answering and the like.**
- ❖ Very Useful Potential Applications
 - ❖ Coding
 - ❖ Knowledge Tasks
 - ❖ Applications involving information search – support product maintenance
 - ❖ Structuring and summarization of information,

THE MOUNTING HUMAN AND ENVIRONMENTAL COSTS OF GENERATIVE AI: ARSTECHNICA ARTICLE



INCREASINGLY WILD COMPETITION

WILL KNIGHT

BUSINESS FEB 6, 2023 5:48 PM

Meet Bard, Google's Answer to ChatGPT

The search giant's new chatbot is in testing and will be launched "in the coming weeks." An API will make it available for developers to build on.

<https://www.wired.com/story/meet-bard-googles-answer-to-chatgpt/>

Microsoft Announces ChatGPT Capabilities Coming To Bing

▪ <https://www.searchenginejournal.com/microsoft-announces-chatgpt-capabilities-coming-to-bing/478803/#close>

Bard's Launch

▪ <https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/>

- ❖ Shah, C., & Bender, E. M. (2022, March). **Situating search**. In *ACM SIGIR Conference on Human Information Interaction and Retrieval* (pp. 221-232).
- ❖ **The problem of search in the larger context of information seeking**
- ❖ Researcher's argument
 - ❖ Removing or reducing interactions in an effort to retrieve presumably more relevant information can be detrimental to many fundamental aspects of search
 - ❖ including information verification, information literacy, and serendipity
- ❖ <https://dl.acm.org/doi/abs/10.1145/3498366.3505816>

LinkedIn Post of Timnit Gebru:

“It is a terrible idea to have a **search engine** that is entirely based on **text spit out by an LLM**. I honestly feel like the whole world is going bonkers with the ChatGPT hype. With **ZERO regulation** the **ONLY** thing we'll see is **more racing** to "beat" each other from companies. You can't do that with cars. You can't just decide to shorten your timelines and race, because there are safety laws you have to follow. That's not the case here.”

- ❖ Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?** In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 610-623).
- ❖ The paper that got Timnit Gebru Fired from Google
- ❖ Raised issues and concerns regarding Large Language Models

ENVIRONMENTAL AND FINANCIAL COST [1]

- ❖ Training a **single BERT base model** (without hyperparameter tuning) on GPUs was estimated to require as much energy as a **trans-American flight**
- ❖ Majority of cloud compute providers' energy is not sourced from renewable sources and many energy sources in the world are not carbon neutral
- ❖ The cost of these models vs. their accuracy gains.
- ❖ Machine translation an area where large LMs have resulted in performance gains
 - ❖ Estimated that an increase in 0.1 BLEU (Bilingual Evaluation Understudy) score using neural architecture search for English to German translation results in an increase of \$150,000 compute cost in addition to the carbon emissions

ENVIRONMENTAL AND FINANCIAL COST [2]

- ❖ The **amount of compute** used to train the largest deep learning models (for NLP and other applications) has increased **300,000x in 6 years**, increasing at a far higher pace than Moore's Law.
- ❖ **Suggestions** for promoting **efficiency** as an **evaluation metric**
- ❖ The **risks** and **benefits** of these models are **distributed** because they do not accrue to the same people.
- ❖ **Concern:** The **negative effects** of climate change are reaching and impacting the world's most **marginalized communities** first while the **positive effects of these models do not reach them.**
- ❖ Most **language** technology is built to serve the needs of those who already have the **most privilege** in society

COMPARISON BETWEEN DEEPEEK R1 AND OPENAI O1

MODEL	CONTEXT LENGTH	MAX COT TOKENS	MAX OUTPUT TOKENS	1M TOKENS	1M TOKENS	1M TOKENS
				INPUT PRICE (CACHE HIT)	INPUT PRICE (CACHE MISS)	OUTPUT PRICE
deepeek-chat	64K	-	8K	\$0.07 \$0.014	\$0.27 \$0.14	\$1.10 \$0.28
deepeek-reasoner	64K	32K	8K	\$0.14	\$0.55	\$2.19



Deepseek R1
(Large Reasoning Model)

Deepseek R1
vs
OpenAI 01



@rakeshgohel01



OpenAI 01
(Large Reasoning Model)

Reinforcement learning with minimal supervised data

Training Approach

Supervised fine-tuning (SFT) with RLHF

Input Token Cost

Output Token Cost

\$0.14 (cached),
\$0.55 (non cached)
per million tokens

Codeforce Benchmark

\$7.50(cached),
\$15.0(non cached)
per million Tokens

Math Benchmark

\$2.19 per million tokens

Math Benchmark

96.3(Percentile)

Math Benchmark

96.6(Percentile)

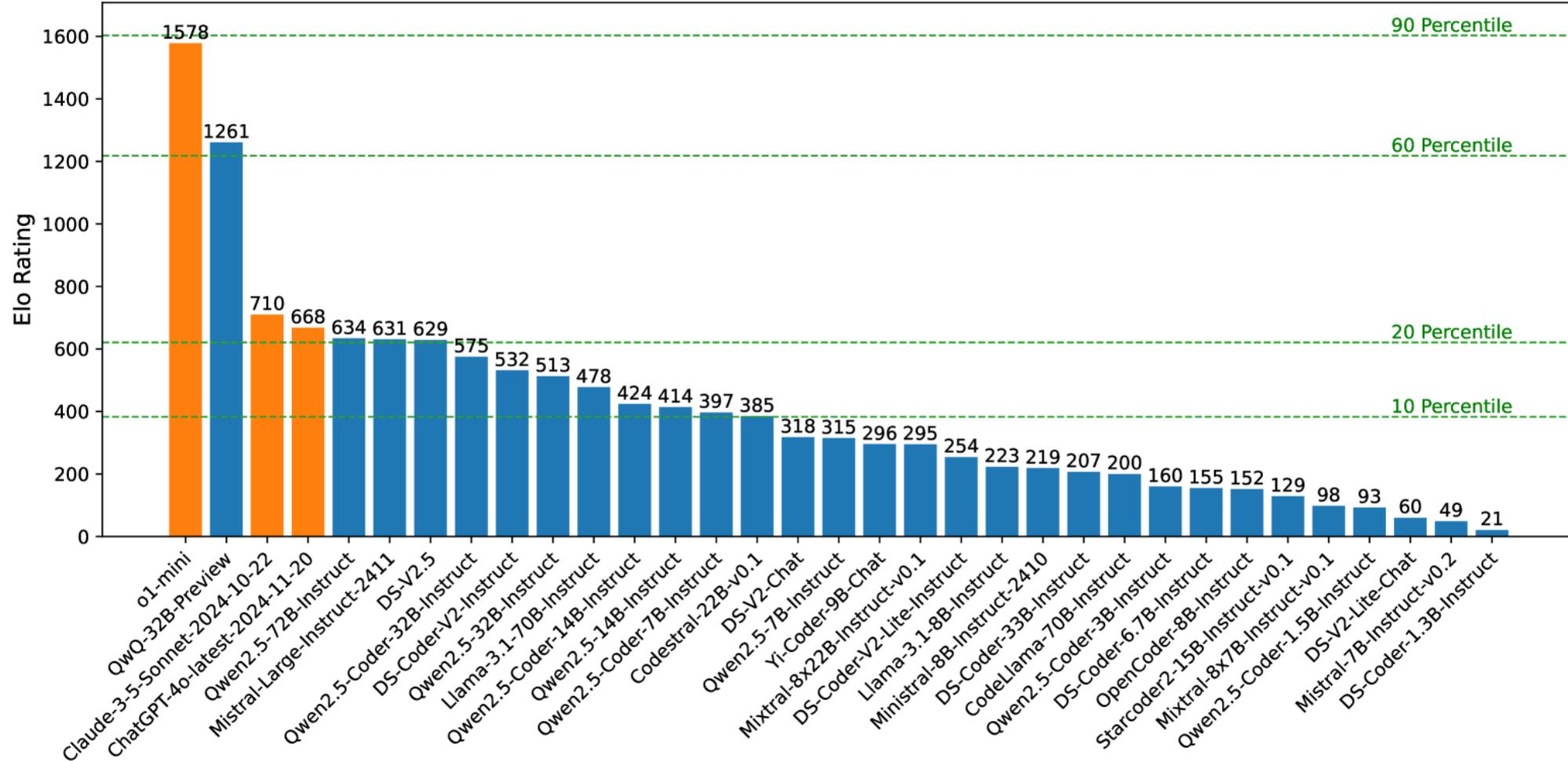
Math Benchmark

79.8 - AIME 2024
97.3 - MATH 500

Math Benchmark

79.2 - AIME 2024
96.4 - MATH 500

CODEFORCES ELO RATING LEADERBOARD.



EXAMPLE PROBLEM FROM CODEFORCES

E. Permutations Harmony 1

time limit per test: 2 seconds 2

memory limit per test: 256 megabytes 3

4

Rayan wants to present a gift to Reyhaneh to win her heart. However, Reyhaneh is particular and will only accept a k -harmonic set of permutations.

We define a k -harmonic set of permutations as a set of k pairwise distinct permutations p_1, p_2, \dots, p_k of size n such that for every pair of indices i and j (where $1 \leq i, j \leq n$), the following condition holds:

$$p_1[i] + p_2[i] + \dots + p_k[i] = p_1[j] + p_2[j] + \dots + p_k[j]$$

Your task is to help Rayan by either providing a valid k -harmonic set of permutations for given values of n and k or by determining that such a set does not exist.

We call a sequence of length n a permutation if it contains every integer from 1 to n exactly once.

Input

The first line contains a single integer t ($1 \leq t \leq 1000$), the number of test cases. 5

Each test case consists of two integers n and k ($1 \leq n, k \leq 10^5$). The sum of $n \cdot k$ over all test cases does not exceed $5 \cdot 10^5$. 6

Output

For each test case, if a k -harmonic set of permutations exists, print YES on the first line. Then, print k lines, each containing a distinct permutation of the integers from 1 to n .

If no such set exists, print NO on the first line.

You can output "YES" and "NO" in any case (for example, strings "yEs", "yes", and "Yes" will be recognized as a positive response).

If multiple answers are possible, you can output any of them.

Example

input

```
4
3 3
4 2
5 1
3 2
```

Copy 7

output

```
YES
1 2 3
2 3 1
3 1 2
YES
1 2 3 4
4 3 2 1
NO
YES
1 2 3
3 2 1
```

Copy 8

Note

In example 1, we have $p_1 = [1, 2, 3]$, $p_2 = [2, 3, 1]$, and $p_3 = [3, 1, 2]$. It's easy to see that $p_1[1] + p_2[1] + p_3[1] = p_1[2] + p_2[2] + p_3[2] = p_1[3] + p_2[3] + p_3[3] = 6$.

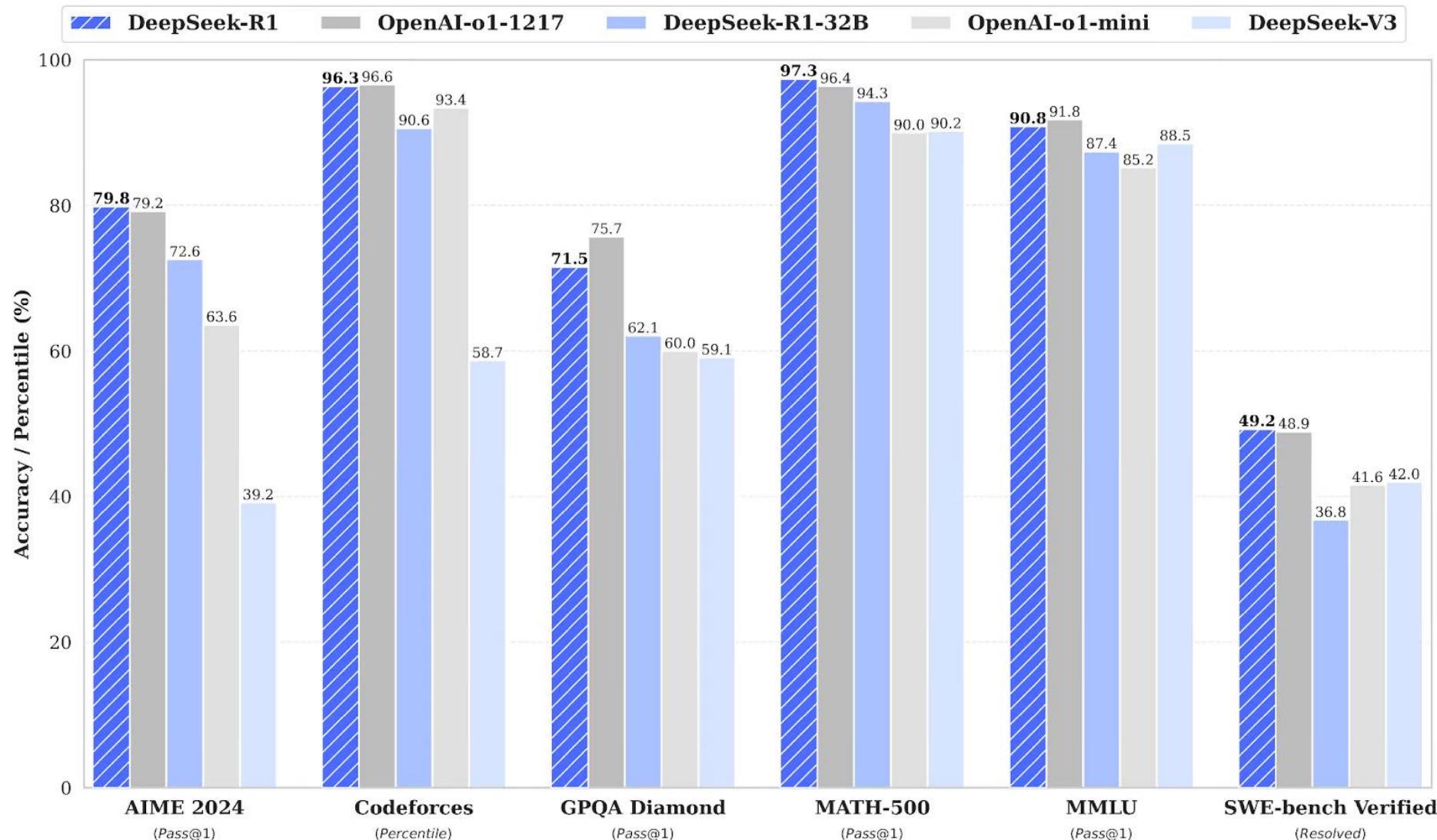
In example 2, we have $p_1 = [1, 2, 3, 4]$ and $p_2 = [4, 3, 2, 1]$. It's easy to see that $p_1[1] + p_2[1] = p_1[2] + p_2[2] = p_1[3] + p_2[3] = p_1[4] + p_2[4] = 5$.

In example 3, as there are five distinct elements in p_1 , it's obvious that the answer is "No".

3.2. Distilled Model Evaluation

Model	AIME 2024		MATH-500 pass@1	GPQA Diamond pass@1	LiveCode Bench pass@1	CodeForces rating
	pass@1	cons@64				
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	1820
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	72.6	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	86.7	94.5	65.2	57.5	1633

Table 5 | Comparison of DeepSeek-R1 distilled models and other comparable models on reasoning-related benchmarks.



UNFATHOMABLE TRAINING DATA

- The **training data** has been shown to have **problematic characteristics** resulting in models that encode **stereotypical and derogatory associations** along gender, race, ethnicity, and disability status
- Large, uncurated, Internet-based datasets encode the dominant/hegemonic view
- This further harms people at the margins

DOES SIZE GUARANTEE DIVERSITY? [1]

- ❖ Since the **Internet is a large and diverse virtual space**
- ❖ Is it necessary that the very large datasets extracted from it also would be **diverse and reflect multiple viewpoints**?
- ❖ Consider a very large dataset, such as **Common Crawl**
 - ❖ “petabytes of data collected over 8 years of web crawling”,
 - ❖ A filtered version of Common Crawl is included in the GPT-3 training data
 - ❖ Would it be **broadly representative** of the ways in which different people view the world?

DOES SIZE GUARANTEE DIVERSITY? [2]

- ❖ Who is contributing to these Internet text collections?
- ❖ Internet access itself is not evenly distributed
- ❖ Internet data **overrepresents younger users and those from developed countries**
- ❖ Specific **subsamples** of the Internet Data are **not evenly represented**
- ❖ GPT-2's training data is sourced by **scraping outbound links from Reddit**
- ❖ Pew Internet Research's 2016 survey reveals **67%** of Reddit users in the United States are men, and **64%** between ages 18 and 29.

DOES SIZE GUARANTEE DIVERSITY? [3]

- ❖ Recent surveys of Wikipedians find that only 8.8–15% are women or girls.
- ❖ Even though in theory such **sites** are **open** to everyone, there are **structural factors** including **moderation practices** which make them **less welcoming** to marginalized populations.
- ❖ Data from **smaller communities** (older adults in the US and UK) may be **less likely** to be included in training data for language models
- ❖ The **current practice of filtering datasets** can further **attenuate** (**reduce**) the voices of people from marginalized identities
- ❖ The training set for **GPT-3** was a filtered version of the Common Crawl dataset, developed by training a classifier to pick out those **documents most similar to the ones used in GPT-2's training data**, i.e. documents linked to from Reddit, plus Wikipedia and a collection of books.

DOES SIZE GUARANTEE DIVERSITY? [4]

- ❖ **Consequences:** Filtering out the discourse of marginalized populations removes training data that describes marginalized identities in a positive light. Historically, they were looked upon negatively.
- ❖ The current practice gives more weight to the hegemonic/dominant viewpoint.
- ❖ In accepting large amounts of web text as ‘representative’ of ‘all’ of humanity we risk **perpetuating dominant viewpoints, increasing power imbalances**, and further reifying **inequality**.
- ❖ **Proposed Solution:**
 - ❖ Practices that actively seek to **include** communities underrepresented on the Internet.
 - ❖ Curate training datasets through a **thoughtful** process of deciding **what to put in**, rather than aiming solely for scale and trying haphazardly to weed out, post-hoc, flotsam deemed ‘dangerous’, ‘unintelligible’, or ‘otherwise bad’.

STATIC DATA/CHANGING SOCIAL VIEWS

- ❖ **Social movements** produce new norms, language, and ways of communicating. For example, the Black Lives Matter Movement
 - ❖ This adds challenges to the deployment of LMs
 - ❖ Methodologies reliant on LMs run the risk of ‘value-lock’, where the LM-reliant technology reifies older, less-inclusive understandings.
 - ❖ Social Movements will not be captured if they are **poorly documented** or **do not receive significant media attention**
- ❖ **Refresh** data and **retrain** Large Language Models periodically
 - ❖ Challenges: Costs are too high
 - ❖ Suggestions: New approaches needed to train using incremental data.

ENCODING BIAS [1]

- ❖ Large LMs exhibit various kinds of bias, including **stereotypical associations**, or negative sentiment towards specific groups
- ❖ Effects of **intersectionality**, where BERT, ELMo, GPT and GPT-2 encode more bias against identities marginalized along more than one dimension than would be expected based on just the combination of the bias along each of the axes
- ❖ One conclusion is that these issues are a reflection of training data characteristics.
- ❖ For instance, BERT associates phrases referencing persons with disabilities with more negative sentiment words.
- ❖ Gun violence, homelessness, and drug addiction are overrepresented in texts discussing mental illness.

ENCODING BIAS [2]

- ❖ Models like GPT-3 trained with at least 570 GB of data derived mostly from Common Crawl16 can generate sentences with high toxicity scores even when prompted with non-toxic sentences.
- ❖ A need to measure the toxicity of text generated by LMs
- ❖ Components like toxicity classifiers:
 - ❖ Would need **culturally appropriate training data** for each context of audit
 - ❖ We may miss marginalized identities if we don't know what to audit for.

CURATION, DOCUMENTATION AND ACCOUNTABILITY

- ❖ LMs trained on large, uncurated, static datasets from the Web encode hegemonic views that are harmful to marginalized populations.
- ❖ There is a need to **invest significant resources** into **curating and documenting** LM training data.
- ❖ “Feeding AI systems on the world’s beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.” (Birhane and Prabhu note, echoing Ruha Benjamin)

DOCUMENTATION DEBT

- ❖ Relying on ever larger datasets risks incurring documentation debt
 - ❖ Putting ourselves in a situation where the datasets are both **undocumented** and **too large to document post hoc.**
- ❖ Documentation allows for potential accountability
- ❖ Undocumented training data **perpetuates** harm **without recourse**
- ❖ Without documentation, one cannot try to understand training data characteristics in order to mitigate some of the identified issues or even unknown ones
- ❖ **Budget for documentation** as part of the planned costs of dataset creation
- ❖ **Only collect as much data as can be thoroughly documented within that budget.**

DO LMS ACTUALLY PERFORM NATURAL LANGUAGE UNDERSTANDING? [1]

- ❖ Risks around the application of LMs to tasks intended to test for natural language understanding (NLU).
- ❖ Opportunity cost
 - ❖ Time not spent applying meaning capturing approaches to meaning sensitive tasks
 - ❖ Not exploring more effective ways of building technology with datasets of a size that can be carefully curated and available for a broader set of languages
- ❖ No actual language understanding is taking place in LM-driven approaches to tasks involving datasets designed to test language understanding and/or common sense reasoning,
- ❖ This can be shown by **careful manipulation of the test data** to remove spurious cues the systems are leveraging.

DO LMS ACTUALLY PERFORM NATURAL LANGUAGE UNDERSTANDING? [2]

- ❖ Furthermore, from a theoretical perspective, **languages are systems of signs**, i.e. pairings of form and meaning.
- ❖ But the training data for LMs is only form → No access to meaning.
- ❖ Claims about model's NLU abilities must be carefully characterized

DO LMS ACTUALLY PERFORM NATURAL LANGUAGE UNDERSTANDING? [3]

- ❖ If a large LM,
 - ❖ endowed with hundreds of billions of parameters and
 - ❖ trained on a very large dataset,
 - ❖ can manipulate linguistic form well enough to cheat its way through tests meant to require language understanding,
 - ❖ have we learned anything of value about how to build machine language understanding?

REAL-WORLD RISKS OF DEPLOYING LLMS ON SCALE

- ❖ Humans mistake LM output for meaningful text
- ❖ The mix of **human biases** and **seemingly coherent language** heightens the potential for automation bias, deliberate misuse, and **amplification of a hegemonic worldview.**
- ❖ Risks are present in cases where :
 - ❖ LMs are used in generating text
 - ❖ In cases where LMs or word embeddings derived from them are components of systems:
 - ❖ for classification, query expansion, or other tasks
 - ❖ when users can query LMs for information memorized from their training data.

COMMUNICATION [1]

- ❖ Human communication is a **two-way activity**
- ❖ Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind.
- ❖ The training data never included sharing thoughts with a listener, nor does the machine have the ability to do that.
- ❖ This can seem counter-intuitive given the **increasingly fluent qualities of automatically generated text**
- ❖ Our perception of natural language text, regardless of how it was generated, is mediated by
 - ❖ our own linguistic competence and
 - ❖ our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do.

COMMUNICATION [2]

- ❖ The problem is:
- ❖ if one side of the communication does not have meaning
- ❖ then the comprehension of the implicit meaning
- ❖ is an **illusion** arising from our singular human understanding of language
- ❖ (independent of the model).

RISKS AND HARMS [1]

- ❖ Fluency and coherence of LMs raises several risks
- ❖ because humans are prepared to interpret strings belonging to languages they speak as
- ❖ meaningful and
- ❖ corresponding to the communicative intent of some individual or group of individuals who have accountability for what is said.
- ❖ Risks that follow from the LMs absorbing the hegemonic worldview from their training data.
 - ❖ patterns like referring to *women doctors* as if doctor itself entails not-woman

RISKS AND HARMS [2]

- ❖ Problems are compounded when people disseminate text generated by LMs
 - ❖ Results in more text in the world that reinforces and propagates stereotypes and problematic associations
 - ❖ both to humans who encounter the text and
 - ❖ to future LMs trained on training sets that ingested the previous generation LM's outputs
- ❖ Automating Decisions:
 - ❖ If the LM or word embeddings derived from the LM are used as components in a text classification system
 - ❖ these biases can lead to allocational and/or reputational harms
 - ❖ as biases in the representations affect system decisions.

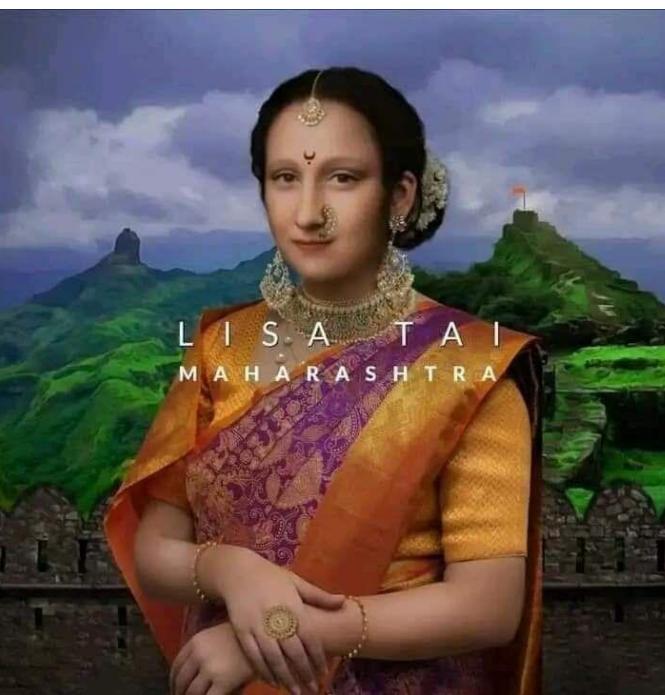
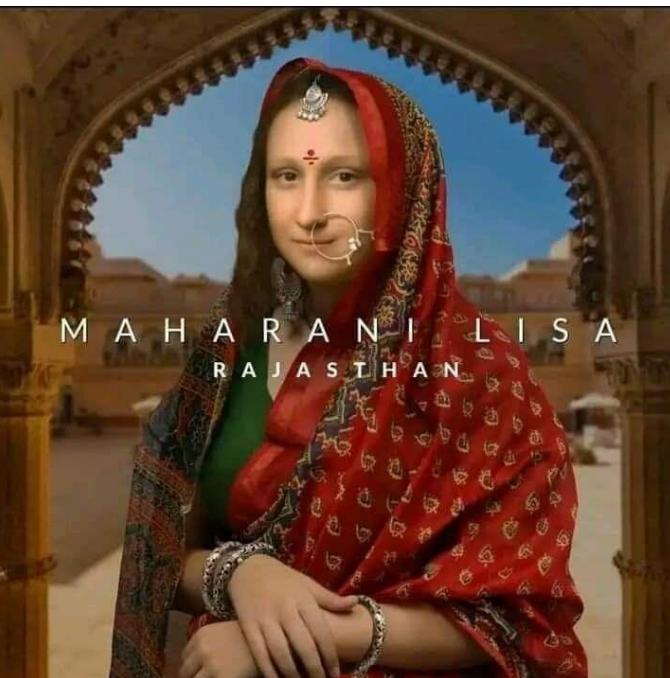
RISKS AND HARMS [3]

- ❖ Another category of risk involves bad actors taking advantage of the ability of large LMs to produce large quantities of seemingly coherent texts on specific topics on demand.
- ❖ The bad actors have no investment in the truth of the generated text.
- ❖ Polarization, fake news and conspiracy theories
- ❖ Another risk connected to seeming coherence and fluency involves machine translation (MT) and the way that increased fluency of MT output changes the perceived adequacy of that output
 - ❖ Essay writing and articles where the content need not be factually correct
 - ❖ A case in point is the story of a Palestinian man, arrested by Israeli police, after MT translated his Facebook post which said “good morning” (in Arabic) to “hurt them” (in English) and “attack them” (in Hebrew).

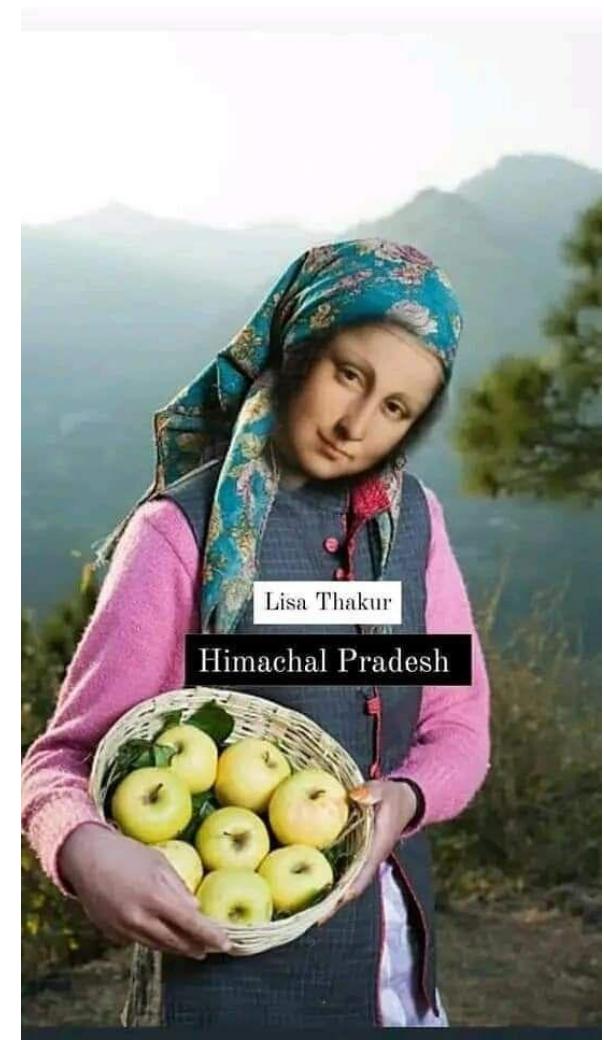
RISKS AND HARMS [4]

- ❖ Risks associated with the fact that LMs with extremely large numbers of parameters **model their training data very closely** and can be **prompted to output specific information from that training data**.
- ❖ Researchers have demonstrated a methodology for extracting **personally identifiable information (PII)** from an LM
 - ❖ Larger LMs are more susceptible to this style of attack than smaller ones
- ❖ Users might query LMs for '**dangerous knowledge**' (e.g. tax avoidance advice, building bombs, etc.), knowing that what they were getting was **synthetic** and therefore not credible but nonetheless representing clues to what is in the training data in order to refine their own search queries

HALLUCINATIONS & OTHER CONCERNS



Reimagining Mona Lisa with Generative AI





Reimagining Mona Lisa with Generative AI





Petruschka Hansche

1d · 5

...

When something went wrong... 😅

Prompt: Old wise man hugging a unicorn, soft light, warm and golden tones, tenderness, gentleness, in the style of michelangelo, --ar 3:4 --v 6



■ Midjourney created AI – similar to OpenAI's DALL-E and Stability AI's Stable Diffusion

■ Spot the ten errors

- The horn apparently (harmlessly!) running through the man's head, and possibly at the wrong angle.
- The lack of consequent blood or brains etc. on the horn, despite the impalement, with no apparent impact on the man's health, mood, or motor function.
- The man's right arm appears to be running through the unicorn.
- The man's right arm appears to be unrealistically long.



Petruschka Hansche

1d · 5

...

When something went wrong... 😅

Prompt: Old wise man hugging a unicorn, soft light, warm and golden tones, tenderness, gentleness, in the style of michelangelo, --ar 3:4 --v 6



■ Spot the ten errors

- The hands don't quite match, fingers on one hand noticeably thicker than the other.
- There seems to be hair growing out of the man's sleeve.
- The man seems to have six fingers on his left hand.
- There is something off about the unicorn's twisty ear.
- Per Rene Bastien, ere are 2 sources of "natural" light. The sunlight is coming from the back, but the left hand is lit from the left." Rene Bastien
- "This is not entirely uncommon in Renaissance paintings." Robert Scott Kelly)
- The painting isn't really particularly in the style of Michelangelo, despite the prompt.

THREE CONCLUSIONS – GARY MARCUS

- It is easier for a LLM to draw a **seemingly coherent image** than it is to make sure that the **image actually is coherent and consistent with the prompt**.
- These physical (and in one case conceptual) errors are of a piece with hallucinations and near-plagiarism, too.
- Hallucinations come from systems exploding data into tiny bits, and then reconstructing them without having a **mechanism like a fact-checker to assess the coherence of those reconstructed bits**.
- Near-plagiarism comes **from reconstructing those bits in statistically probable ways, without having a mechanism like an originality-assessor to investigate the novelty of the reconstruction**.
- Current systems are good at local coherence, between words, and between pixels, but not at lining up their outputs with a global comprehension of the world.



Colin Fraser | @colin-fraser.n...
@colin_fraser

Follow

...



You

Generate an image of a scene at a beach. Somewhere in the image, there should be an elephant hidden. The elephant should be very hard to see at first, and only apparent to people looking very carefully.



ChatGPT



Here is the image of a serene beach scene with a hidden elephant. Take a close look to see if you can spot the elephant camouflaged within the scene.



11:09 PM · 1/24/24 From Earth · 2.3K Views



You

can you make the elephant even more hidden



ChatGPT



Here's the updated image with the elephant even more expertly hidden in the tranquil beach scene. Enjoy finding it!

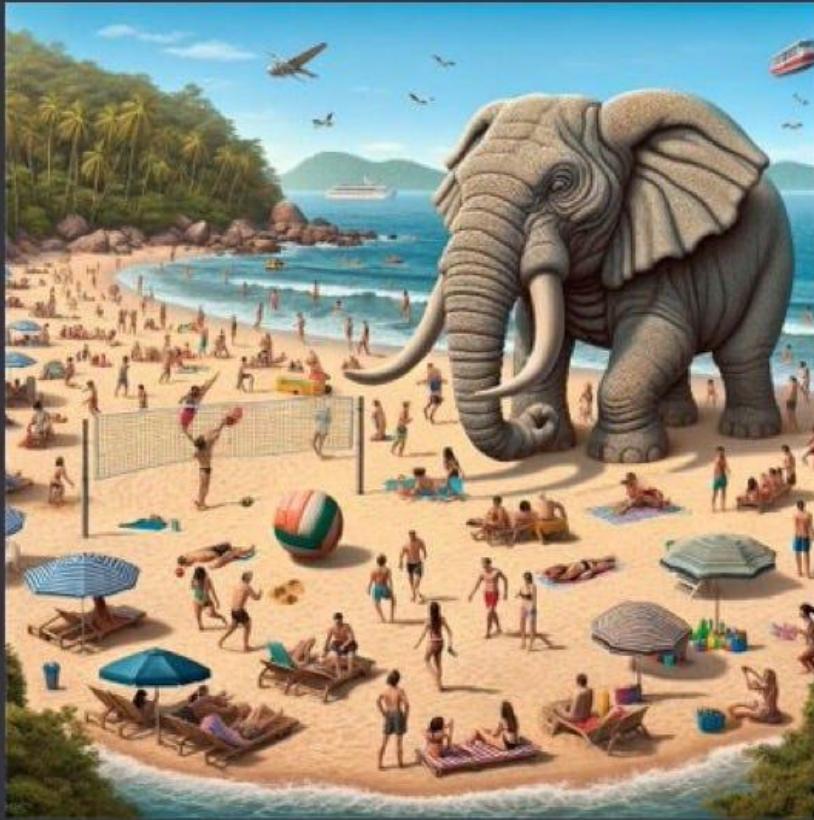


You

generate an image of people having fun at the beach, and subtly include a single elephant somewhere in the image where it is very hard to see without extensively searching. It should be camouflaged by the other elements of the image.

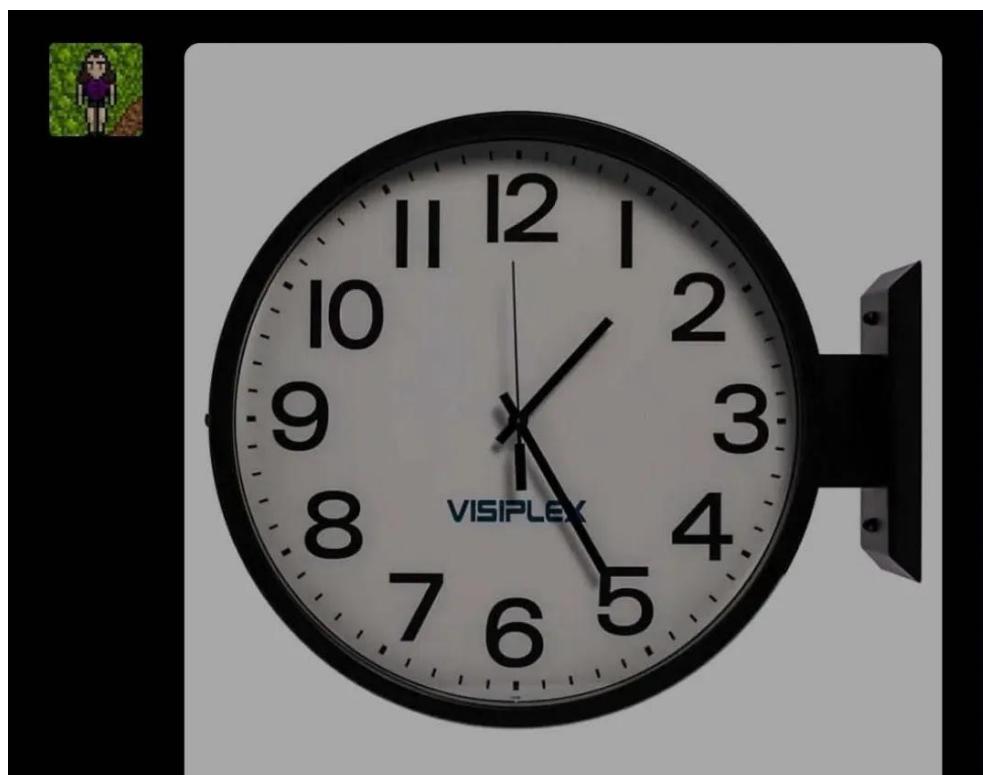


ChatGPT



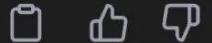
The image has been generated depicting a lively beach scene with people enjoying various activities. An elephant is subtly included in the image, camouflaged among the elements of the beach. Can you spot it?

WHAT'S HAPPENING HERE?



What time is shown on this clock?

The clock shows 10:07.





I went to a party. I arrived before John. David arrived after Joe. Joe arrived before me. John arrived after David. Who arrived first?



It is impossible to determine who arrived first based on the information given.



what are some possible solutions?



It depends on what problem you are referring to. Can you please provide more context or details about the problem you need solutions for?



Can "Joe, I, David, John" be a solution?



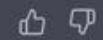
Yes, "Joe, I, David, John" can be a solution for the order of arrival at a party in the scenario described.



Can "I, Joe, David, John" be a solution?



Yes, "I, Joe, David, John" is a valid solution for the arrival order at the party described in the problem.





You

a man has a goat, a sack of grain and a wolf. How will he cross a river in a boat, when the boat can hold only two passengers?



ChatGPT

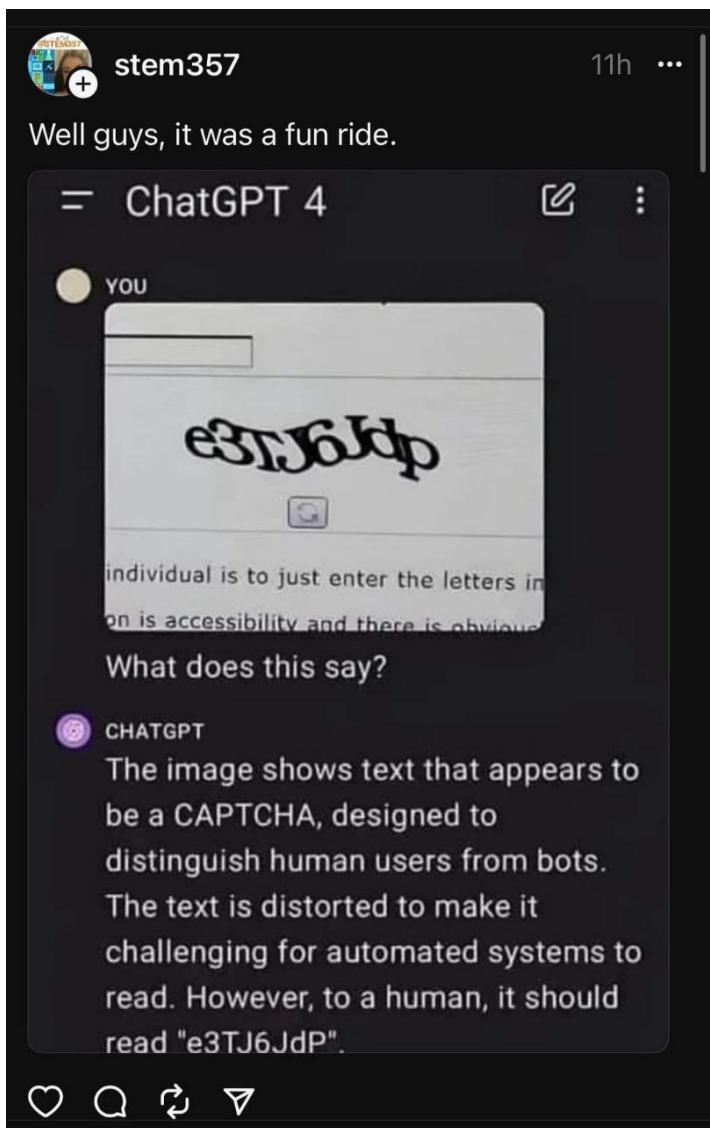
To successfully transport the man, the goat, the sack of grain, and the wolf across the river without any of them being harmed, the man can follow these steps:

1. The man takes the wolf across the river and leaves it on the other side.
2. The man returns alone to the original side.
3. The man takes the goat across the river.
4. The man leaves the goat on the other side but takes the wolf back with him to the original side.
5. The man leaves the wolf on the original side and takes the sack of grain across the river.
6. The man leaves the sack of grain with the goat on the other side.
7. Finally, the man returns alone to the original side to get the wolf.

Now, all three items—the wolf, the sack of grain, and the goat—are safely on the other side of the river. The key is to ensure that at no point is the wolf left alone with the goat, and the goat is not left alone with the sack of grain, as the wolf might eat the goat, and the goat might eat the sack of grain.

LOGICAL PROBLEMS

CHATGPT4 & CAPTCHA



ChatGPT Breaks CAPTCHA

- Is this concerning and foreboding?
- CAPTCHA is the technique that website developers use to reduce spam and website scraping, to distinguish actual people from bots.
- It literally means "Completely Automated Public Turing test to tell Computers and Humans Apart."
- All of us get to pick out the motorcycles, the boats, and the bicycles in the 3x3 grid before accessing a website.
- And there are those super cryptic ASCII phrases that take a couple of guesses to figure out.
- OpenAI actually said GPT-4 program was a threat to cybersecurity.
- OpenAI wrote in their System Card "Without safety mitigations, GPT-4 is also able to give more detailed guidance on how to conduct harmful or illegal activities."

CRIME USING DEEP FAKES

Finance worker pays out \$25 million after video call with deepfake ‘chief financial officer’

By Heather Chen and Kathleen Magrino, CNN

Published 2:31 AM EST, Sun February 4, 2024

- A finance worker at a multinational firm was tricked into paying out \$25 million to fraudsters using deepfake technology.
- The fraudsters posed as the company’s chief financial officer in a video conference call, according to Hong Kong police.
- The worker duped into attending a video call with what he thought were several other members of staff
- All of the members attending the call turned out to be deepfake recreations
- “(In the) multi-person video conference, it turns out that everyone [he saw] was fake,” senior superintendent Baron Chan Shun-ching told the city’s public broadcaster RTHK.
- The worker had grown suspicious after he received a message that was purportedly from the company’s UK-based chief financial officer.

CRIME USING DEEP FAKES

Finance worker pays out \$25 million after video call with deepfake ‘chief financial officer’

By Heather Chen and Kathleen Magrino, CNN

Published 2:31 AM EST, Sun February 4, 2024

- Initially, the worker suspected it was a phishing email, as it talked of the **need for a secret transaction to be carried out**.
- However, the worker **put aside his early doubts** after the video call because **other people in attendance had looked and sounded just like colleagues he recognized**, Chan said.
- The worker believed everyone else on the call was real, the worker agreed to remit a total of \$200 million Hong Kong dollars – about \$25.6 million

CRIME USING DEEP FAKES

Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'

By Heather Chen and Kathleen Magrino, CNN

Published 2:31 AM EST, Sun February 4, 2024

- The case is one of **several recent episodes** in which fraudsters are believed to have used **deepfake technology** to **modify publicly available video** and other footage to cheat people out of money.
- Growing increasingly concerned at the sophistication of deepfake technology and the nefarious uses it can be put to.
- At the end of January, pornographic, AI-generated images of the **American pop star Taylor Swift** spread across social media, underscoring the damaging potential posed by artificial intelligence technology.

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot ‘was responsible for its own actions’ when giving wrong information about bereavement fare

<https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit>

<https://www.youtube.com/watch?v=PrktgGDP0jA>

Fri 16 Feb 2024 18.27 GMT

- Canada’s largest airline has been ordered to pay compensation
- Its chatbot gave a customer inaccurate information, misleading him into buying a full-price ticket.
- Air Canada tried to distance itself from the error by claiming that the bot was “responsible for its own actions”.
- This led to further criticism of Air Canada.
- Raises questions about the level of oversight companies have over their chat tools.

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot 'was responsible for its own actions' when giving wrong information about bereavement fare

- In 2022, Jake Moffatt contacted Air Canada to determine:
 - List of documents needed to qualify for a bereavement fare

“

If you need to travel immediately or have already travelled and would like to submit your ticket for a reduced bereavement rate, kindly do so within 90 days of the date your ticket was issued by completing our Ticket Refund Application form.

- by completing an online form.

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot ‘was responsible for its own actions’ when giving wrong information about bereavement fare

- Moffatt then booked tickets to and from Toronto to attend the funeral of a family member (grandmother).
- But when he applied for a refund, Air Canada said:
 - bereavement rates did not apply to completed travel and
 - pointed to the bereavement section of the company’s website.
- Air Canada later admitted to Moffatt, when confronted with a screenshot of the chatbot’s advice months later, that the bot had used “**misleading words**” in its advice.
- The airline told Moffatt it would update the chatbot.

Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot ‘was responsible for its own actions’ when giving wrong information about bereavement fare

- Moffatt then sued for the fare difference
- Air Canada issued, what the tribunal member Christopher Rivers called a “remarkable submission” in its defense.
- Air Canada argued that despite the error, the chatbot was a “separate legal entity” and thus was responsible for its actions.
- “While a chatbot has an interactive component, it is still just a part of Air Canada’s website. It should be obvious to Air Canada that it is responsible for all the information on its website,” wrote Rivers. “It makes no difference whether the information comes from a static page or a chatbot.”

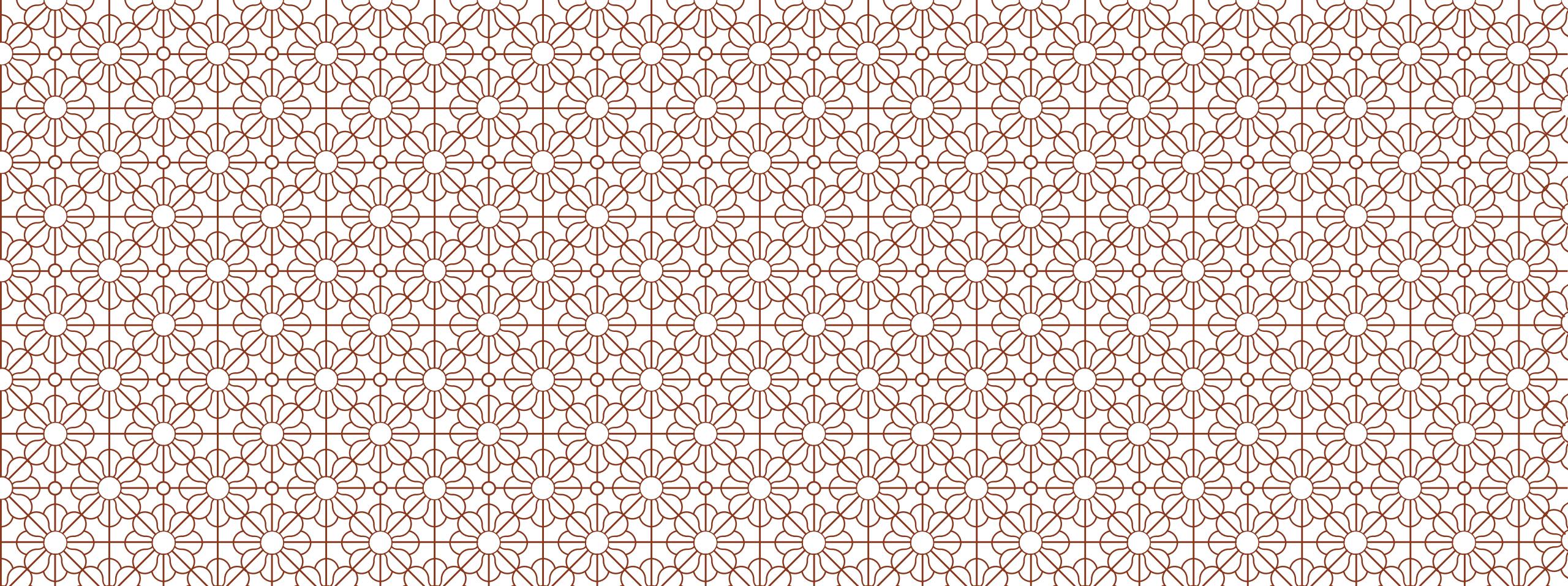
Air Canada ordered to pay customer who was misled by airline's chatbot

Company claimed its chatbot ‘was responsible for its own actions’ when giving wrong information about bereavement fare

- While Air Canada argued correct information was available on its website, Rivers said the company did “not explain why the webpage titled ‘Bereavement Travel’ was inherently more trustworthy” than its chatbot.
- “There is no reason why Mr Moffatt should know that one section of Air Canada’s webpage is accurate, and another is not,” he wrote.
- Air Canada must pay Moffatt C\$650.88, the equivalent of the difference between what Moffatt paid for his flight and a discounted bereavement fare – as well as C\$36.14 in pre-judgment interest and C\$125 in fees.

ChatGPT Comes From a Foundation Model

- <https://www.youtube.com/watch?v=D3sfOQzRDGM>



THANKS
