香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

CSC6052/5051/4100/DDA6307/ MDS5110 Natural Language Processing

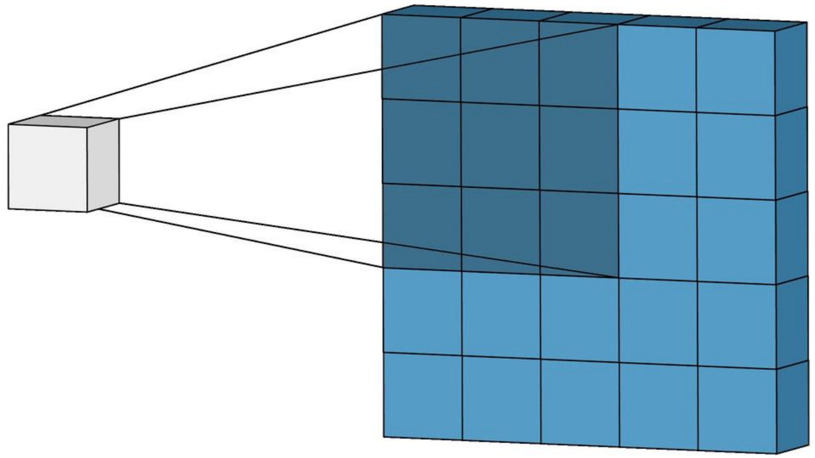
Lecture 5-1: Large Language Models (LLMs)

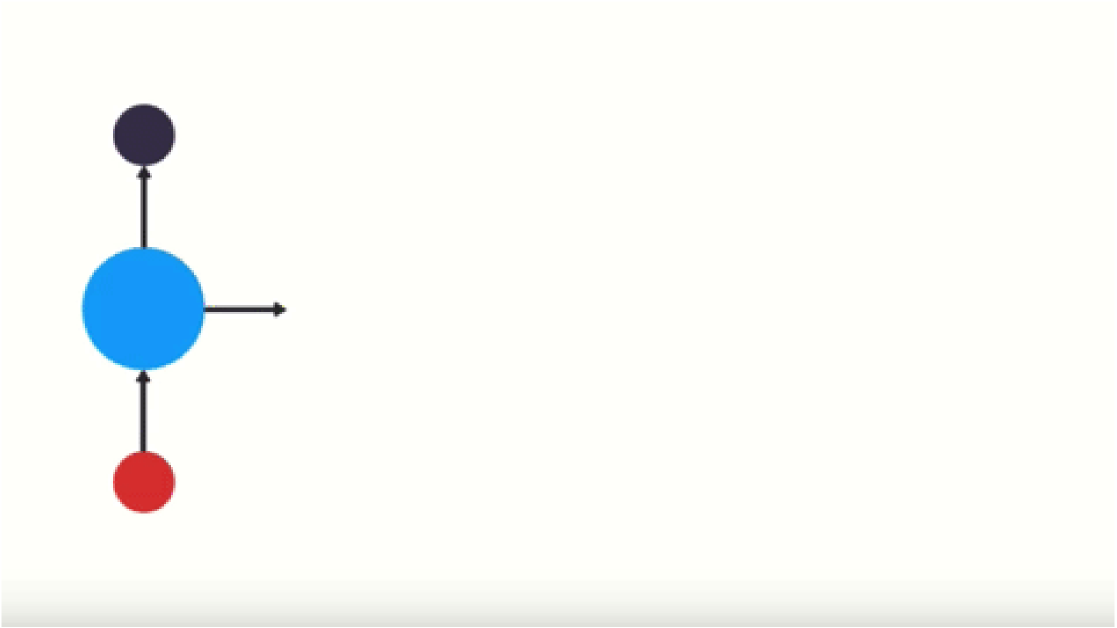
Spring 2025
Benyou Wang
School of Data Science

To recap and an overview

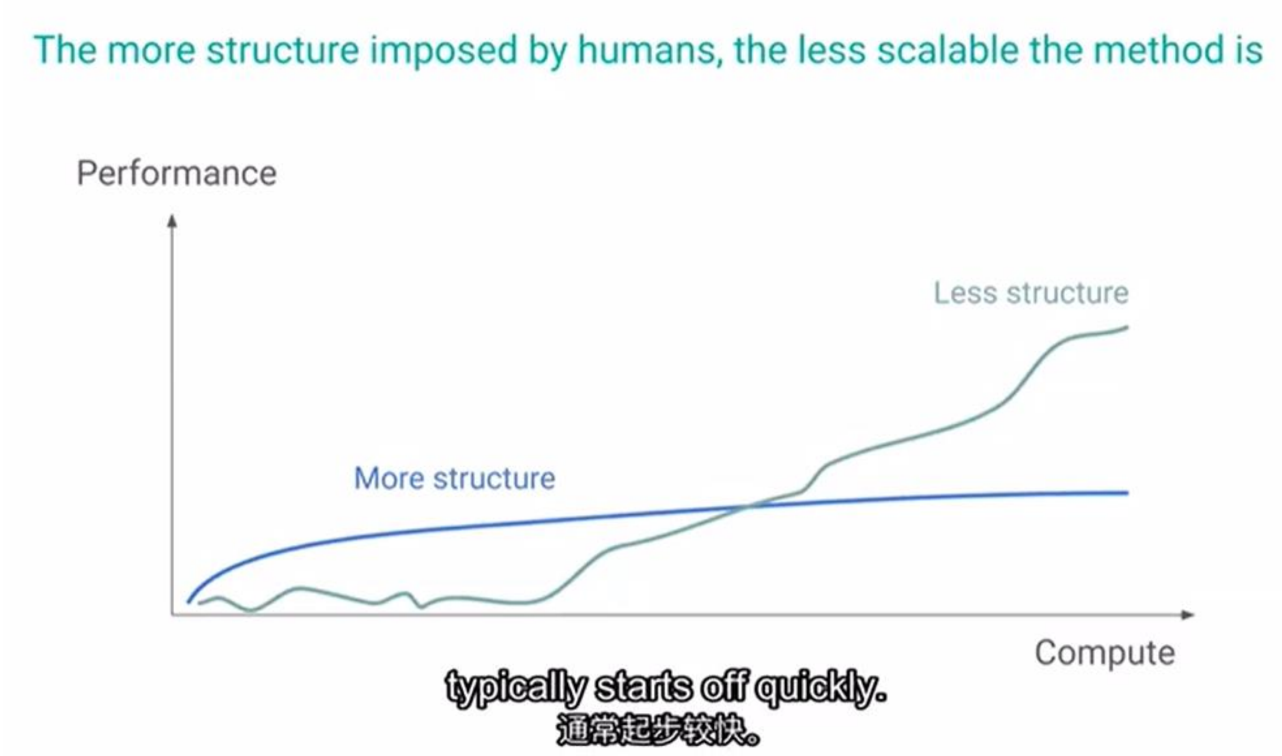
Inductive bias of composition

CNN: **local** composition within a window
RNN: **recurrently** compose tokens from left to right or right to left.





A video you must watch

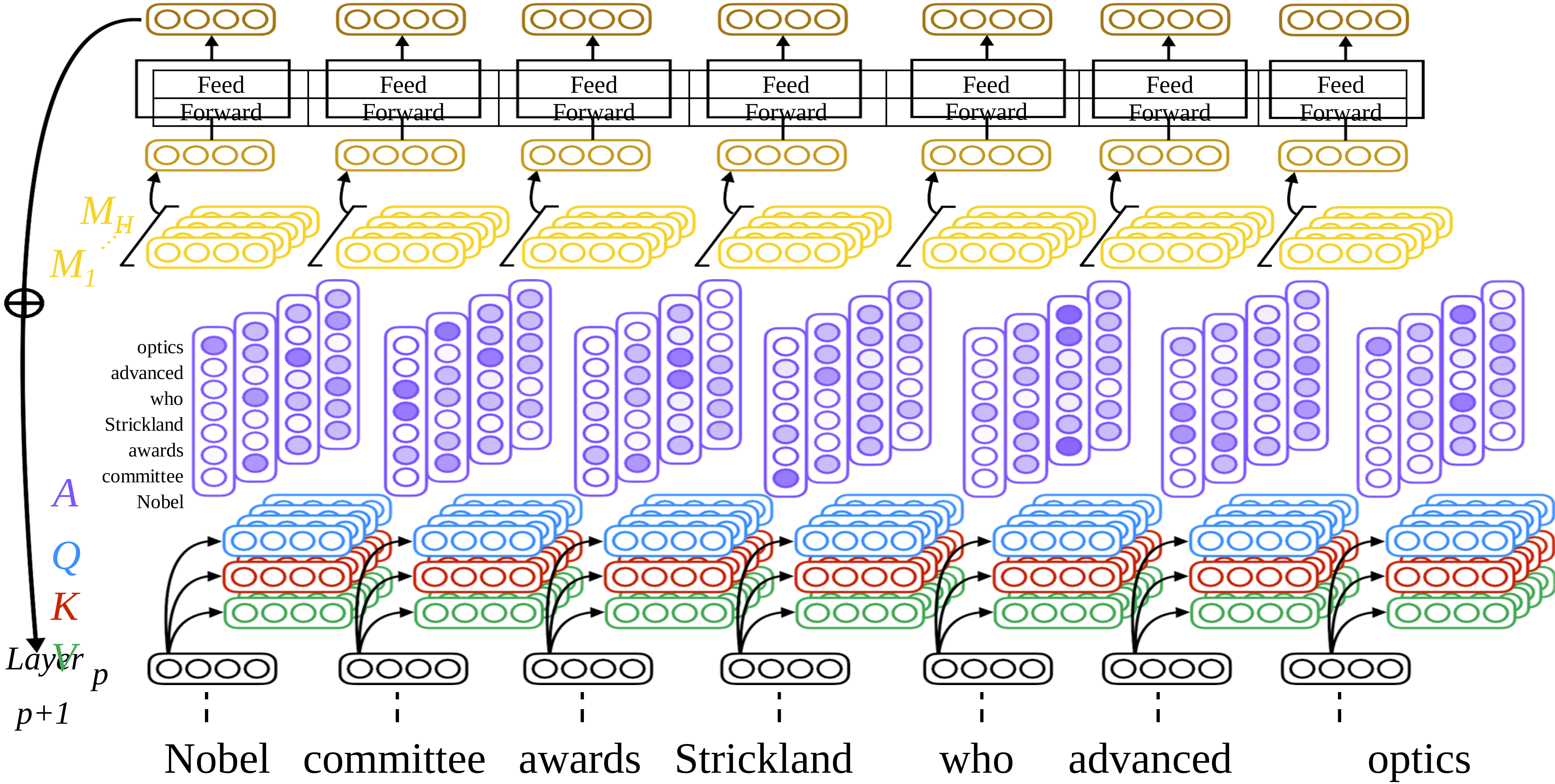


Reducing inductive bias (local or recurrent bias) and take **full attention**!

https://www.youtube.com/watch?v=kYWUEV_e2ss

Multi-head self-attention

[Vaswani et al. 2017]



Scaling Law

Performance depends strongly on scale! We keep getting better performance as we scale the model, data, and compute up!

Scaling Laws for Neural Language Models

Jared Kaplan *
Johns Hopkins University, OpenAI
jaredk@jhu.edu

Sam McCandlish*
OpenAI
sam@openai.com

Tom Henighan
OpenAI
henighan@openai.com

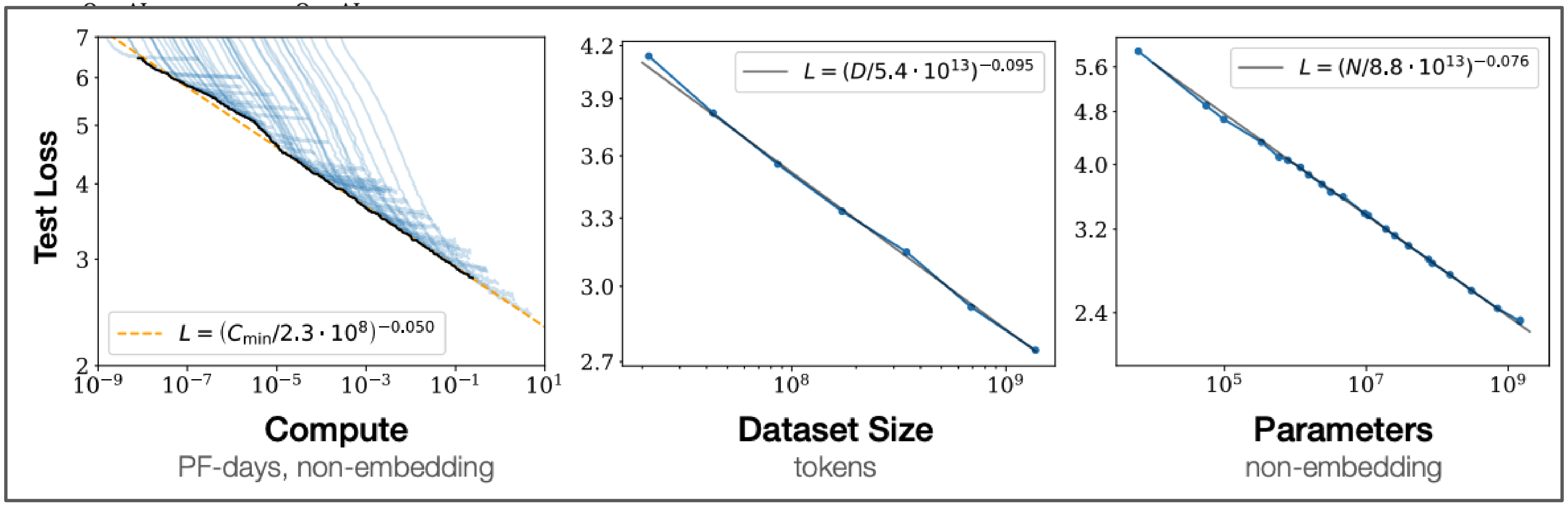
Tom B. Brown
OpenAI
tom@openai.com

Benjamin Chess

Rewon Child

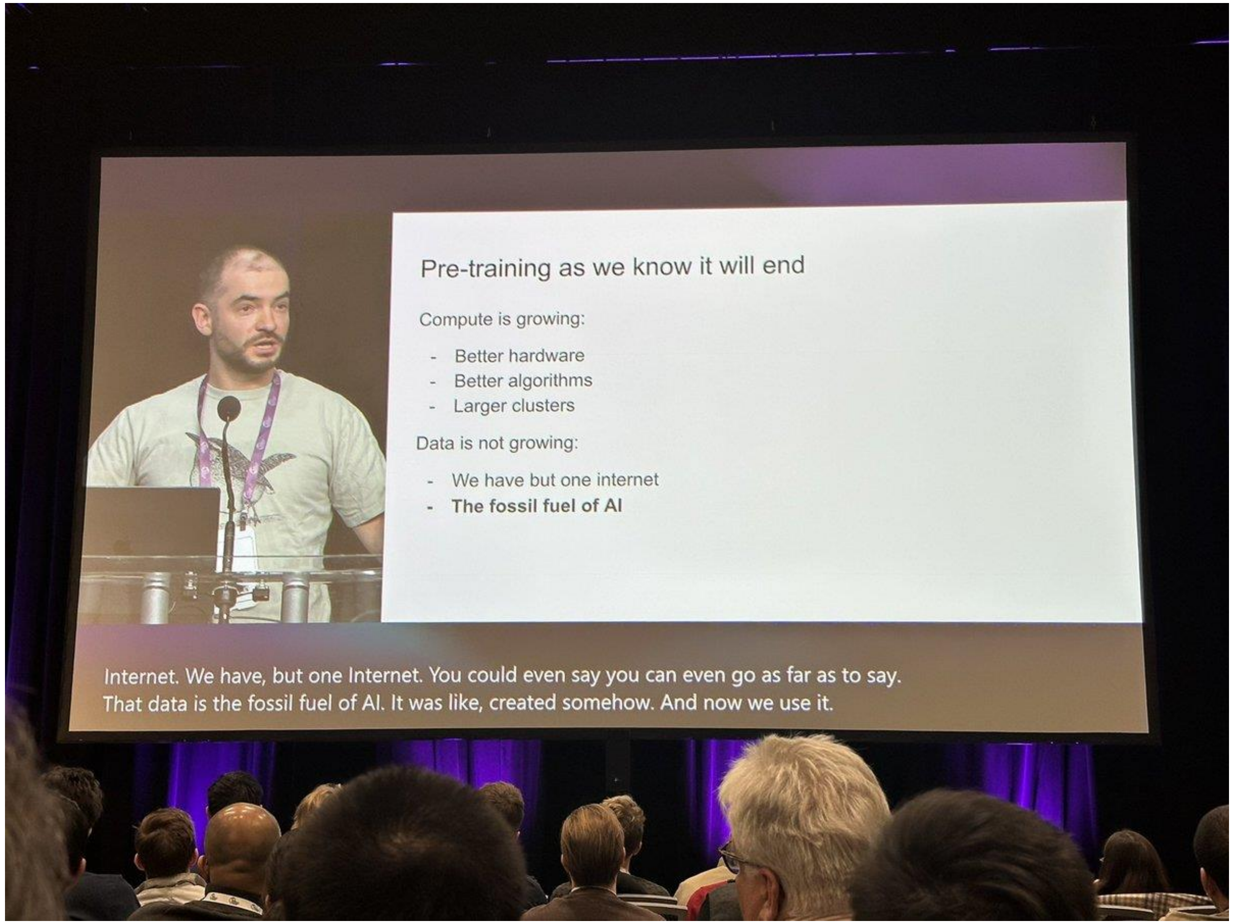
Scott Gray
OpenAI
scott@openai.com

Alec Radford
OpenAI
alec@openai.com



Emergent abilities of large language models (TMLR ‘22).
J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, & W. Fedus.

Ilya Sutskever says **scaling (pretraining) will ends**

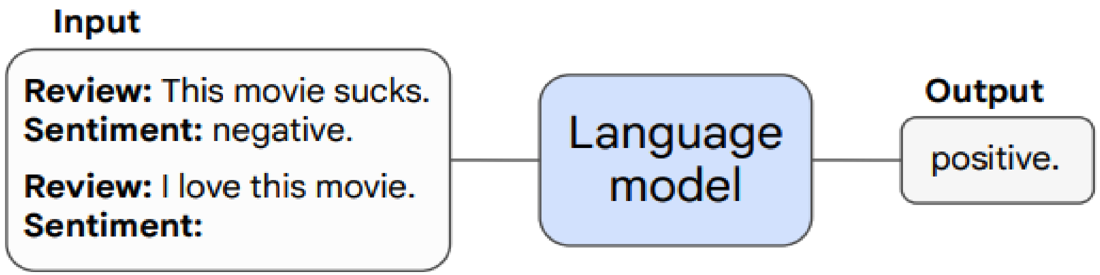


https://youtu.be/1yvBqasHLZs

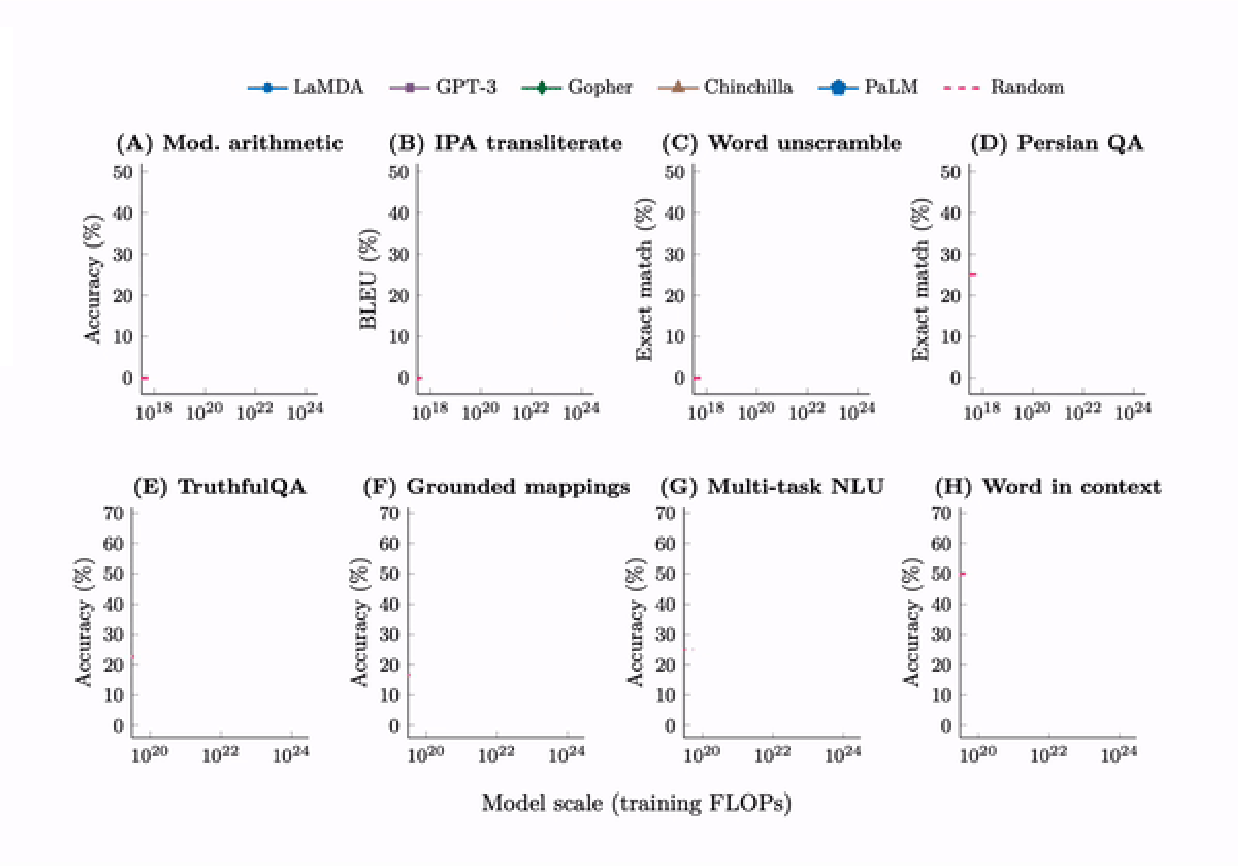
Emergent properties in LLMs:

Some ability of LM is not present in smaller models but is present in larger models

Emergent Capability: Few-shot prompting

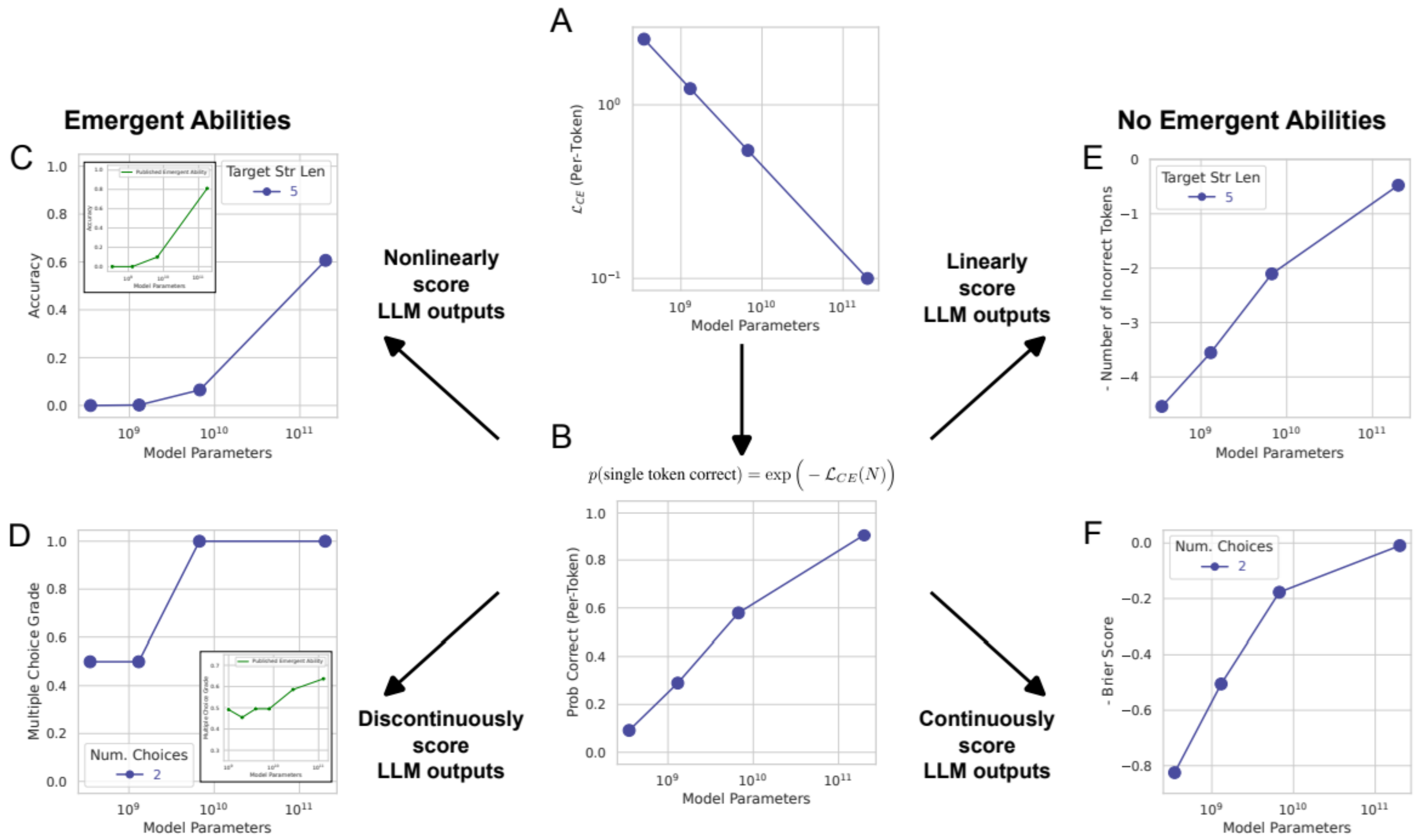


> A few-shot prompted task is emergent if it achieves random accuracy for small models and above-random accuracy for large models.



https://docs.google.com/presentation/d/1yzbmYB5E7G8lY2-KzhmArmPYwwl7o7CUST1xRZDUu1Y/edit?resourcekey=0-6_TnUMoK WCk_FN2BiPxmbw#slide=id.g1fc34b3ac18_0_27

Emergent capabilities may be a consequence of metric choice



It seems that emergent ability of a model only occurs if the measure of per-token error rate of any model is scaled **non-linearly or discontinuously**.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are Emergent Abilities of Large Language Models a Mirage? https://browse.arxiv.org/pdf/2304.15004.pdf

Outline

1. What are LLMs
2. How Large is Large LMs?

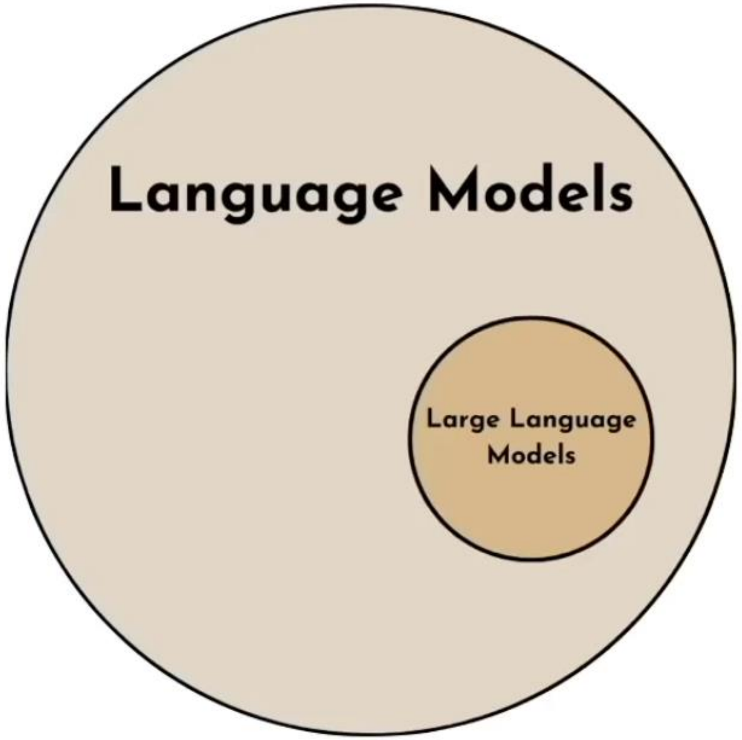
What are LLMs

Think about that:

What is the difference between Large Language models (LLMs) and language models?

Just larger?

LM vs. LLM



Quantitatively

Number of model parameters
i.e. ~10-100 Billion

Qualitatively

Emergent properties[1]
i.e. Zero-shot learning

Zero-shot Learning in LLMs

The capability of a (machine learning) model to complete a task it was not explicitly trained to do

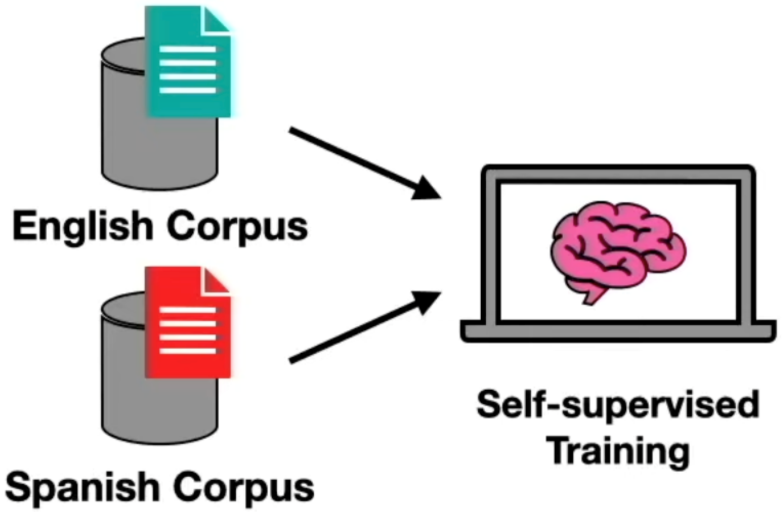
Old Way
(Supervised learning)

Train model on 1k-1M labelled examples

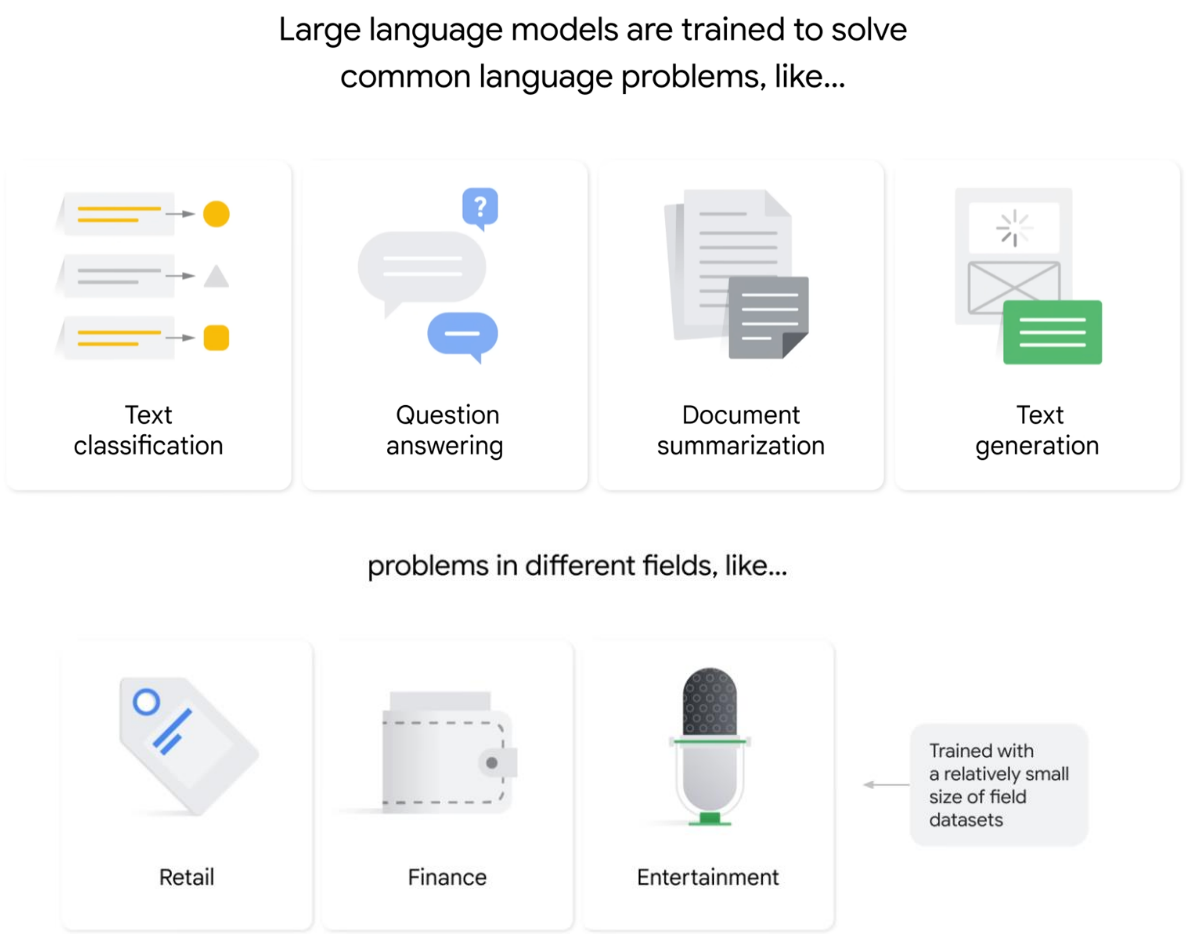
Input	Label
Hello	English
Hola	Spanish
How's it going?	English
...	...
Esta Bien	Spanish

New Way
(Self-supervised learning)

Train (very large) model on (very large) corpus



What is an LLM （one model for nearly everything）



Also called “foundation model” [1]

[1] Rishi Bommasani et al. On the Opportunities and Risks of Foundation Models. https://arxiv.org/pdf/2108.07258.pdf

DL hypothesis

Anything a human do in **0.1 seconds**, a big 10-layer neural network can do, too.

At than moment we could not training a much lager model.

Jason Wei’ Rule of thumb

language models can do (with decent accuracy) most things that an average human can do in **1 minute**.

https://ai4comm.media.mit.edu/slides/emergence.pdf

AGI

Artificial general intelligence (AGI) refers to the hypothetical intelligence of a machine that possesses the ability to understand or learn **any intellectual task that a human being can**.

Do you believe that LLMs could achieve it?

A possible way to AGI

LLM + Agent
- LLM is the **brain**
- An agent framework equip AGI with **tools**

Brain: think longer (o1-like thinking)
Tools: equip it with external knowledge/information or rules

Qiguang Chen, Libo Qin, Jiaqi Wang, Jinxuan Zhou, Wanxiang Che. Unlocking the Capabilities of Thought: A Reasoning Boundary Framework to Quantify and Optimize Chain-of-Thought.

The direction for AGI?

Now people is working on complex reasoning (LLM brain is not that clever enough).

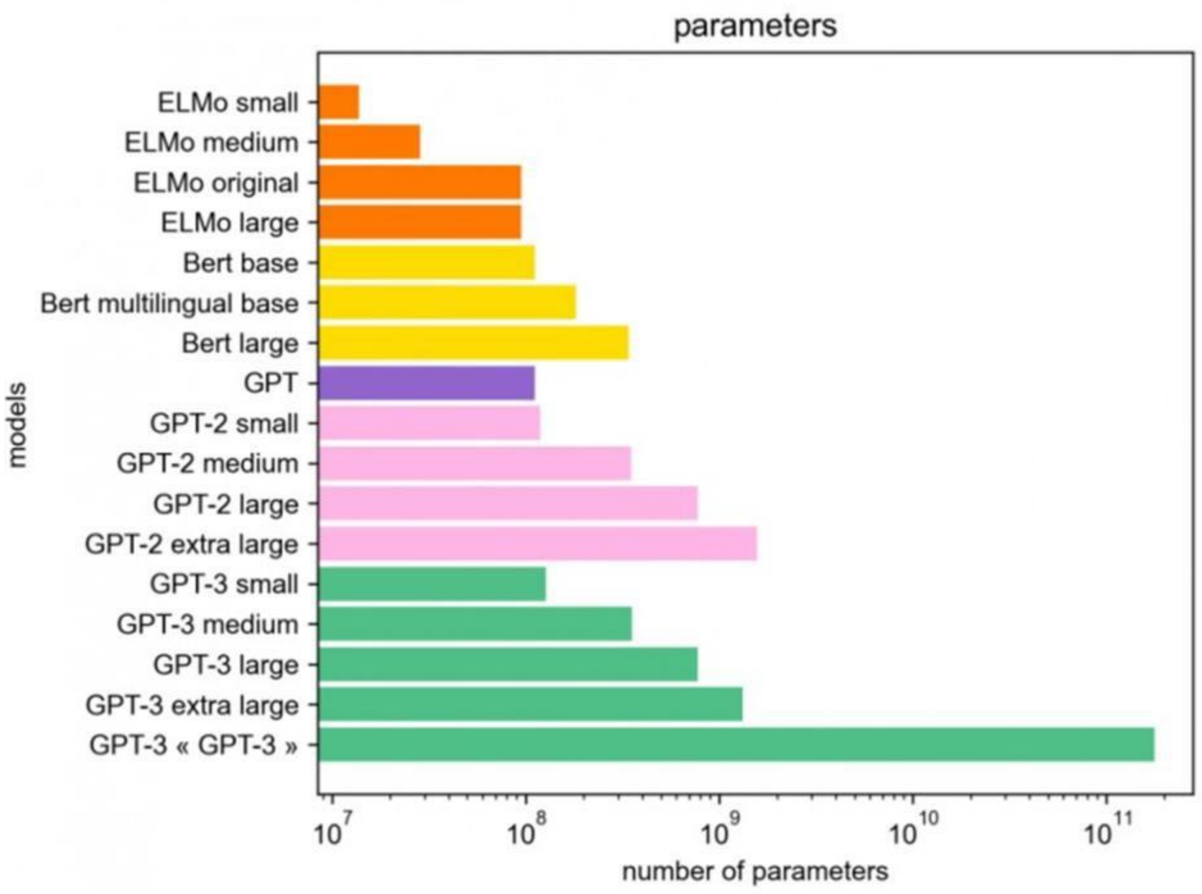
The current answer is **RL based on output reward**?

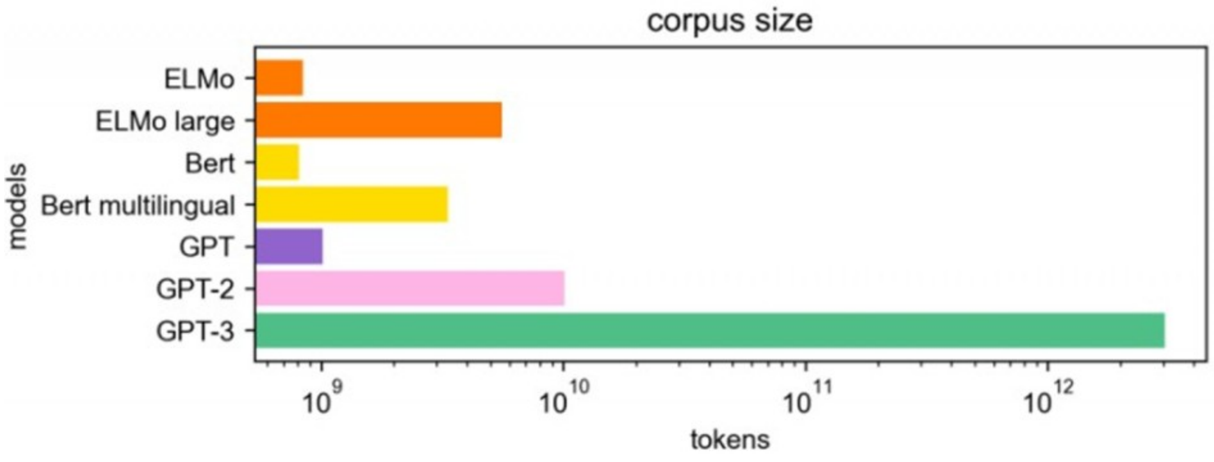
If RL is done and the performance becomes saturated. An open question (neural-symbolic) will be:

whether **formal** languages helps process verifications?

How large is large?

How Large are "Large" LMs?





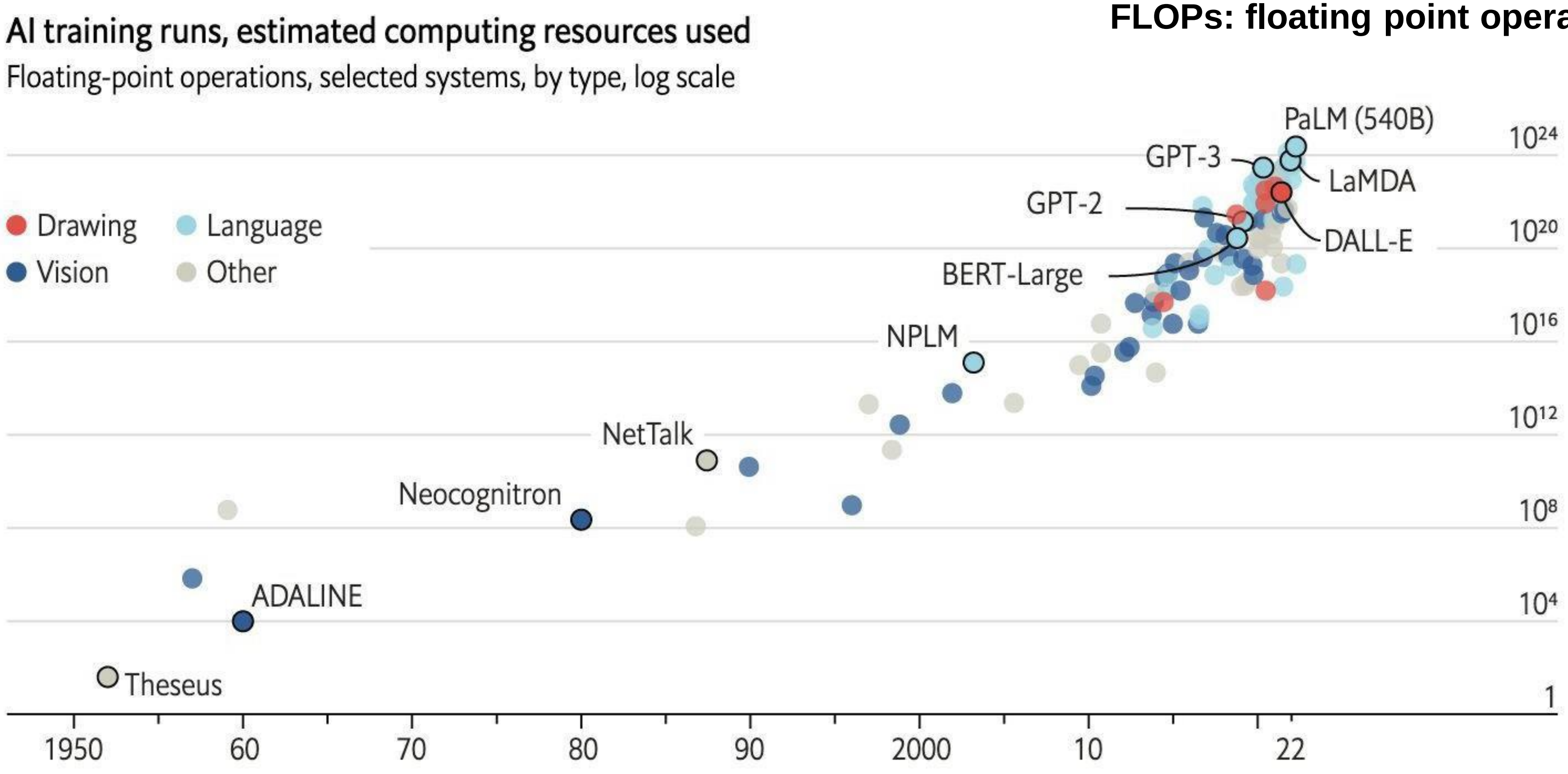
More recent models: PaLM (540B), OPT (175B), BLOOM (176B)…

https://huggingface.co/blog/large-language-models

Image source: https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/

Large Language Models - yottaFlops of Compute

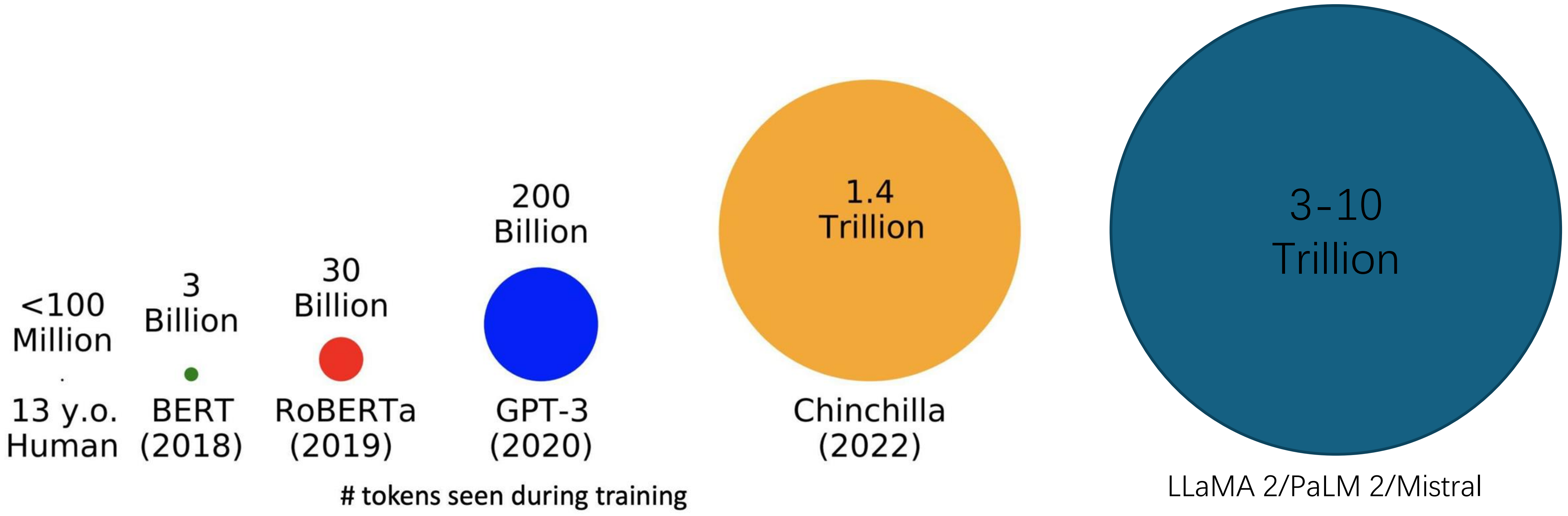
FLOPs: floating point operations



~~GPT 4: with 1.8T parameters (equivalent to 280B dense parameter) -- it is said!~~

https://web.stanford.edu/class/cs224n/slides/cs224n-2023-lecture11-prompting-rlhf.pdf

Large Language Models - Hundreds of Billions of Tokens

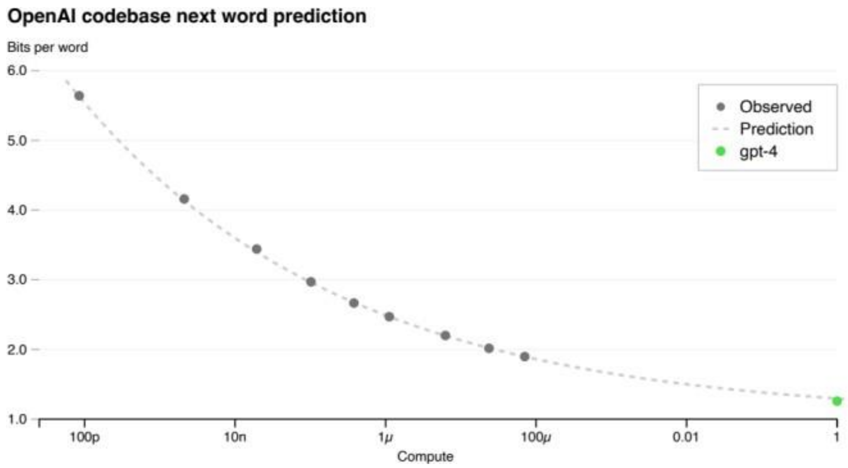


~~GPT 4: with 13T tokens -- it is said!~~

https://babylm.github.io/

Some basics for large language models

- Scalable network **architecture** (Transformer vs. CNN/RNN)
- Scalable **objective** (**conditional/**auto-regressive LM vs. Masked LM)



- Scalable **data** (plain texts are everywhere vs. supervised data)
  - The fossil fuel (data) seems over.

OpenAI, GPT-4 Technical Report, https://cdn.openai.com/papers/gpt-4.pdf

How large is “large”?

- ❖ In BERT era
  - ➢ Base models: BERT/RoBERTa (100M),
  - ➢ Large one: 300M

- ❖ T5 era
  - ➢ Base models: 200M
  - ➢ small models: 60M
  - ➢ Large: 770M
  - ➢ Much lager: 3B and 11B (XXXL)

- ❖ LLM
  - ➢ Base models: probably 7B to 13B
  - ➢ Small models: 60M

Interestingly, small language model becomes popular

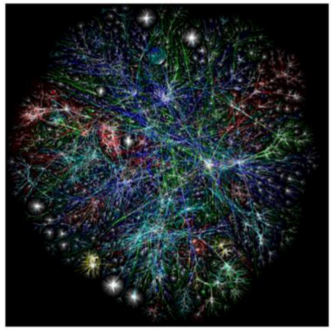
TinyLLaMA: 1.1B
MobileVLM: 1.4B and 2.7B
MobiLlama 0.5B
MobileLLM: 0.1B and 0.3B

[1] MobileLLM: Optimizing Sub-billion Parameter Language Models for On-Device Use Cases. https://arxiv.org/pdf/2402.14905.pdf
[2] MobiLlama: Towards Accurate and Lightweight Fully Transparent GPT. https://arxiv.org/abs/2402.16840
[3] MobileVLM : A Fast, Strong and Open Vision Language Assistant for Mobile Devices https://arxiv.org/abs/2312.16886.
[4] TinyLlama: An Open-Source Small Language Model. https://arxiv.org/abs/2401.02385

Why LLMs?

Why LLMs: Learning/intelligence as compression

Think of it like compressing the internet.

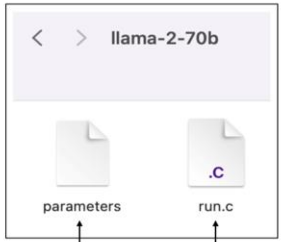


Chunk of the internet,
~10TB of text

6,000 GPUs for 12 days, ~$2M
~1e24 FLOPS

parameters.zip

~140GB file



140GB

~500 lines
of C code

MacBook

Next word prediction forces the neural network to learn a lot about the world:

Ruth Marianna Handler (née **Mosko**; November 4, 1916 – April 27, 2002) was an American businesswoman and inventor. She is best known for inventing the Barbie doll in 1959,[2] and being co-founder of toy manufacturer Mattel with her husband Elliot, as well as serving as the company's first president from 1945 to 1975.[3]

The Handlers were forced to resign from Mattel in 1975 after the Securities and Exchange Commission investigated the company for falsifying financial documents.[3][4]

Early life [edit]

Ruth Marianna Mosko[5][2][3] was born on November 4, 1916, in Denver, Colorado, to Polish-Jewish immigrants Jacob Moskowicz, a blacksmith, and Ida Moskowicz, née Rubenstein.[6]

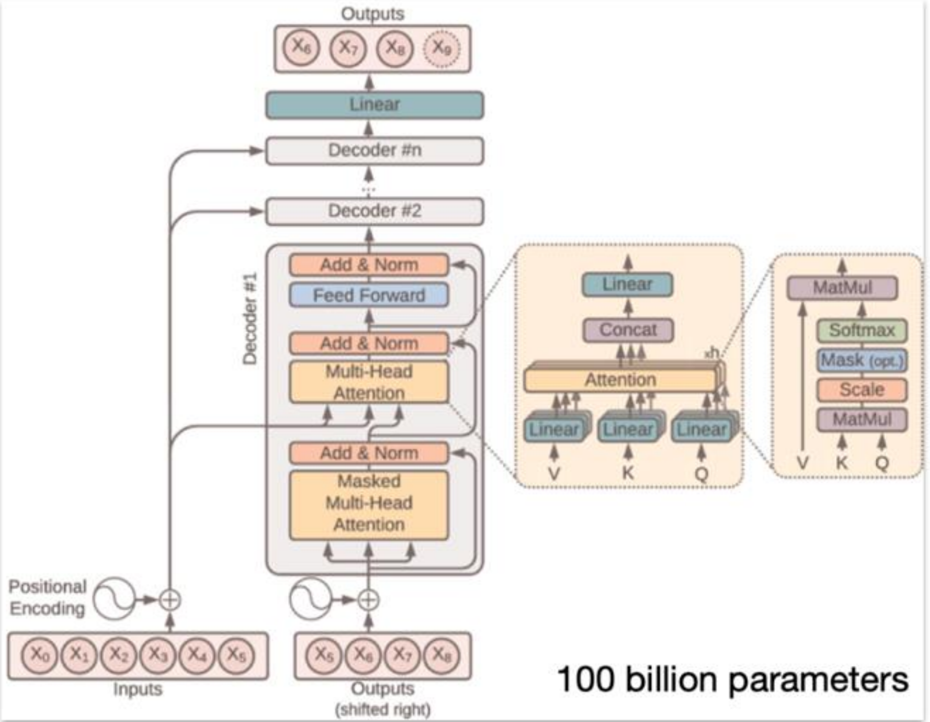
She married her high school boyfriend, Elliot Handler, and moved to Los Angeles in 1938, where she found work at Paramount.[7]

Ruth Handler

Handler in 1961

Born	Ruth Marianna Mosko November 4, 1916 Denver, Colorado, U.S.
Died	April 27, 2002 (aged 85)[1] Los Angeles, California, U.S.

Why does it Work?



Little is known in full detail…

- Billions of parameters are dispersed through the network
- We know how to iteratively adjust them to make it better at predictior
- We can measure that this works, but we don't really know how the billions of parameters collaborate to do it.

They build and maintain some kind of knowledge database, but it is a bit strange and imperfect:

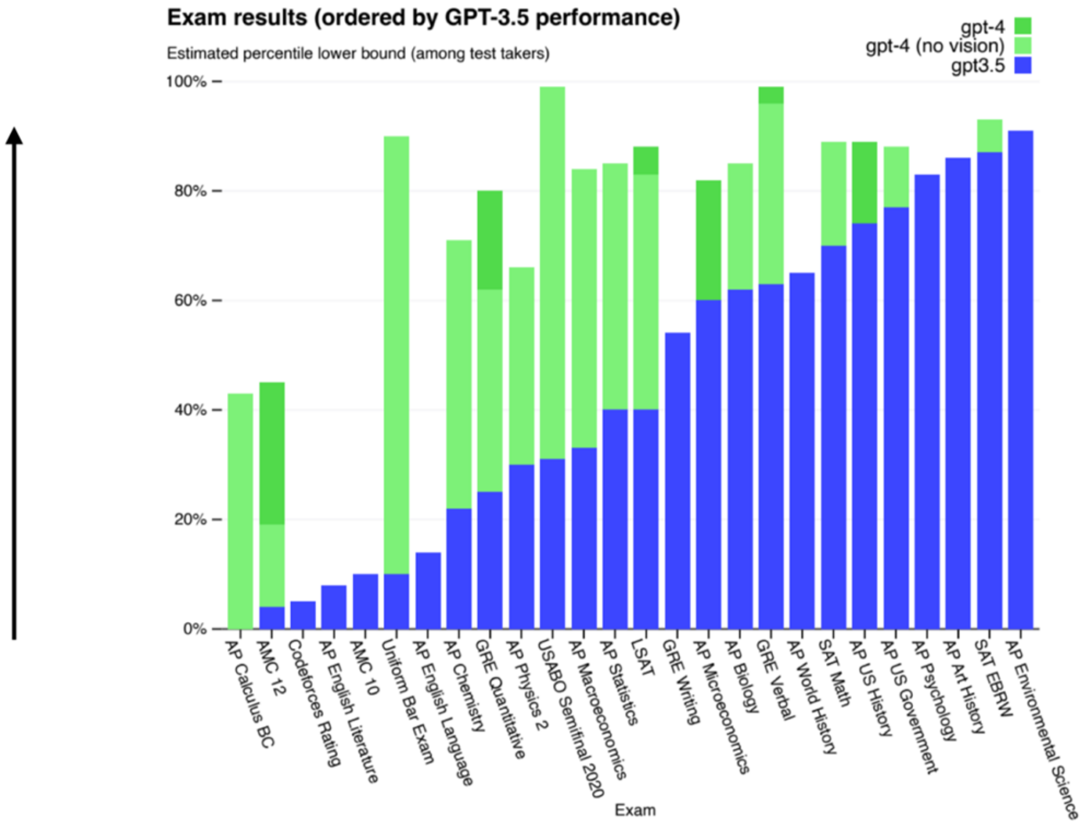
Recent viral example: "reversal curse"

Q: "Who is Tom Cruise's mother"?
A: Mary Lee Pfeiffer ✅

Q: "Who is Mary Lee Pfeiffer's son?"
A: I don't know ❌ 🤔

=> think of LLMs as mostly inscrutable artifacts, develop correspondingly sophisticated evaluations.

Benefit of being larger, e.g., GPT-4 over ChatGPT 3.5



[Sparks of Artificial General Intelligence: Early experiments with GPT-4, Bubuck et al. 2023]

We can expect a lot more “general capability” across all areas of knowledge:

Why Larger language models

- More world **knowledge** (LAMA)
  - Language models as knowledge base?
- Larger capacity to learn problem-solving **Abilities**
  - Coding, revising articles, reasoning etc.
- Better **generalization** to unseen tasks

- **Emergent** ability (涌现能力)

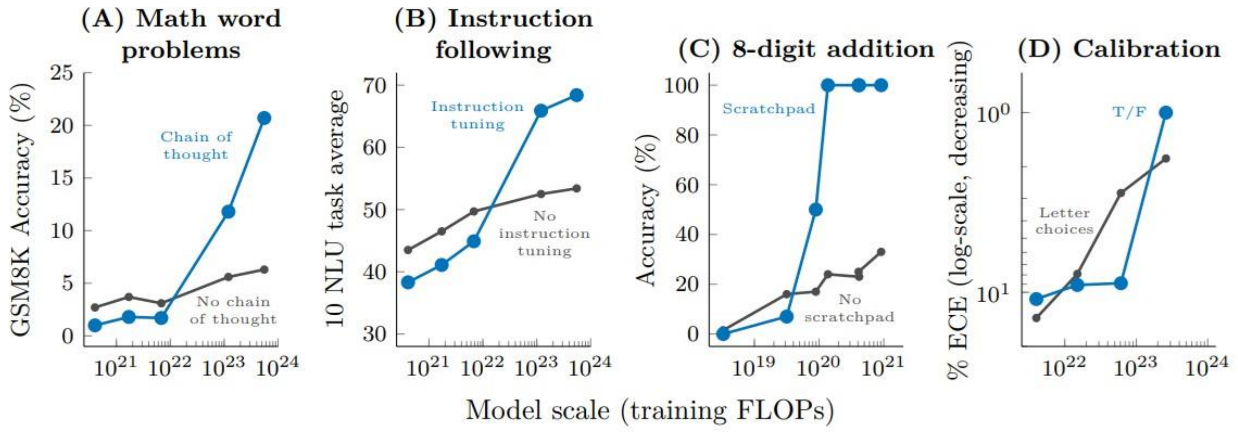


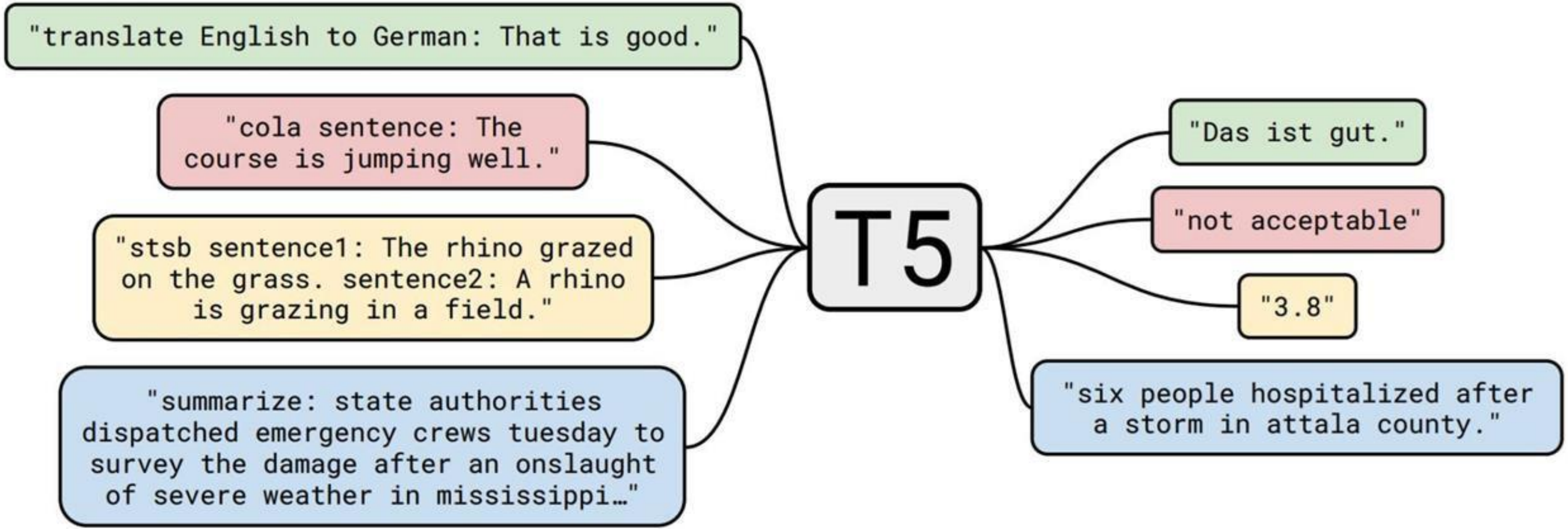
Figure 3: Specialized prompting or finetuning methods can be emergent in that they do not have a positive effect until a certain model scale. A: Wei et al. (2022b). B: Wei et al. (2022a). C: Nye et al. (2021). D: Kadavath et al. (2022). An analogous figure with number of parameters on the x-axis instead of training FLOPs is given in Figure 12. The model shown in A-C is LaMDA (Thoppilan et al., 2022), and the model shown in D is from Anthropic.

Jared Kaplan et. al **Scaling Laws for Neural Language Models**
Jason Wei et. Al. Emergent Abilities of Large Language Models.

Why LLMs?

Generalization :

One single model to solve many NLP tasks



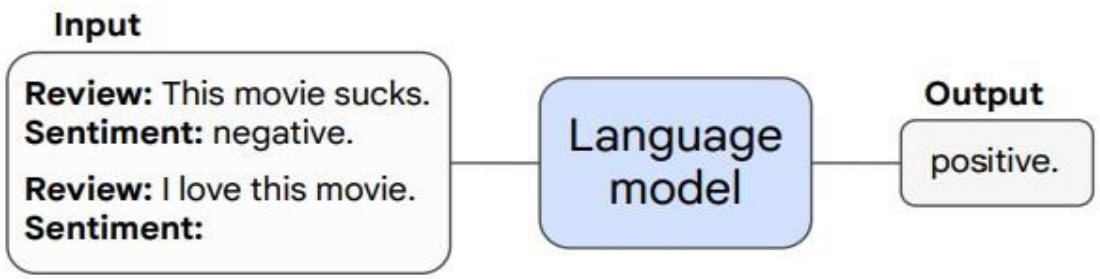
It could even generalizes to new tasks, following the phylosity of FLAN

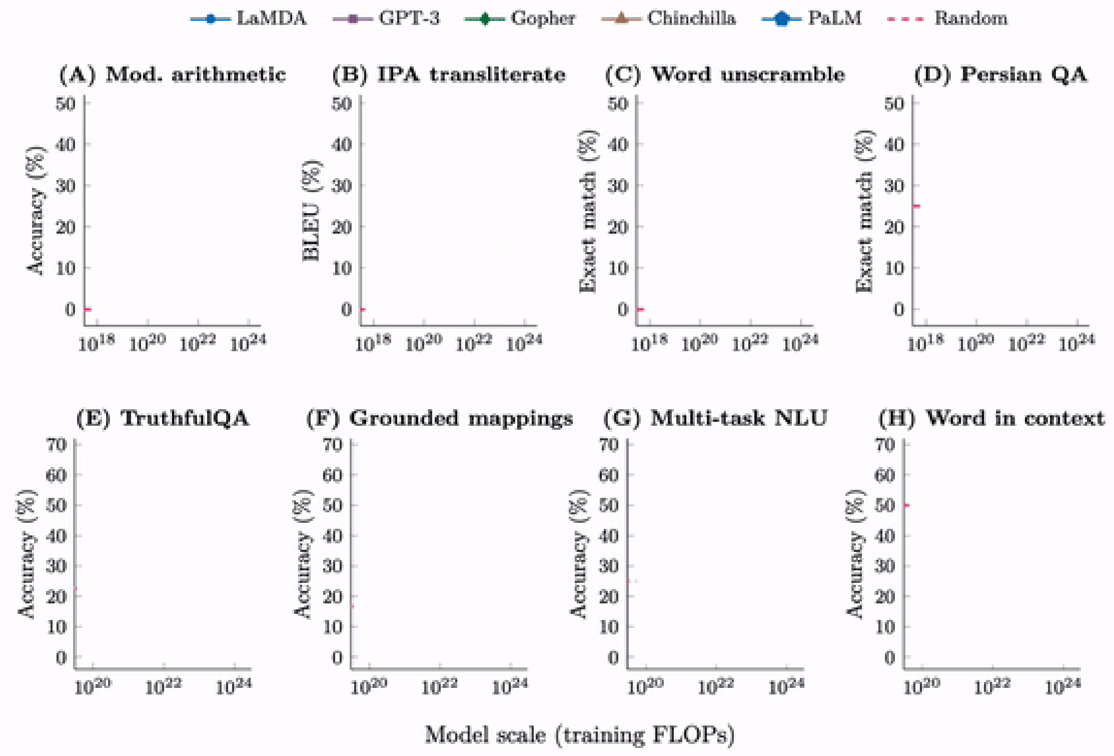
Why LLMs?

Emergent properties in LLMs:

Some ability of LM is not present in smaller models but is present in larger models

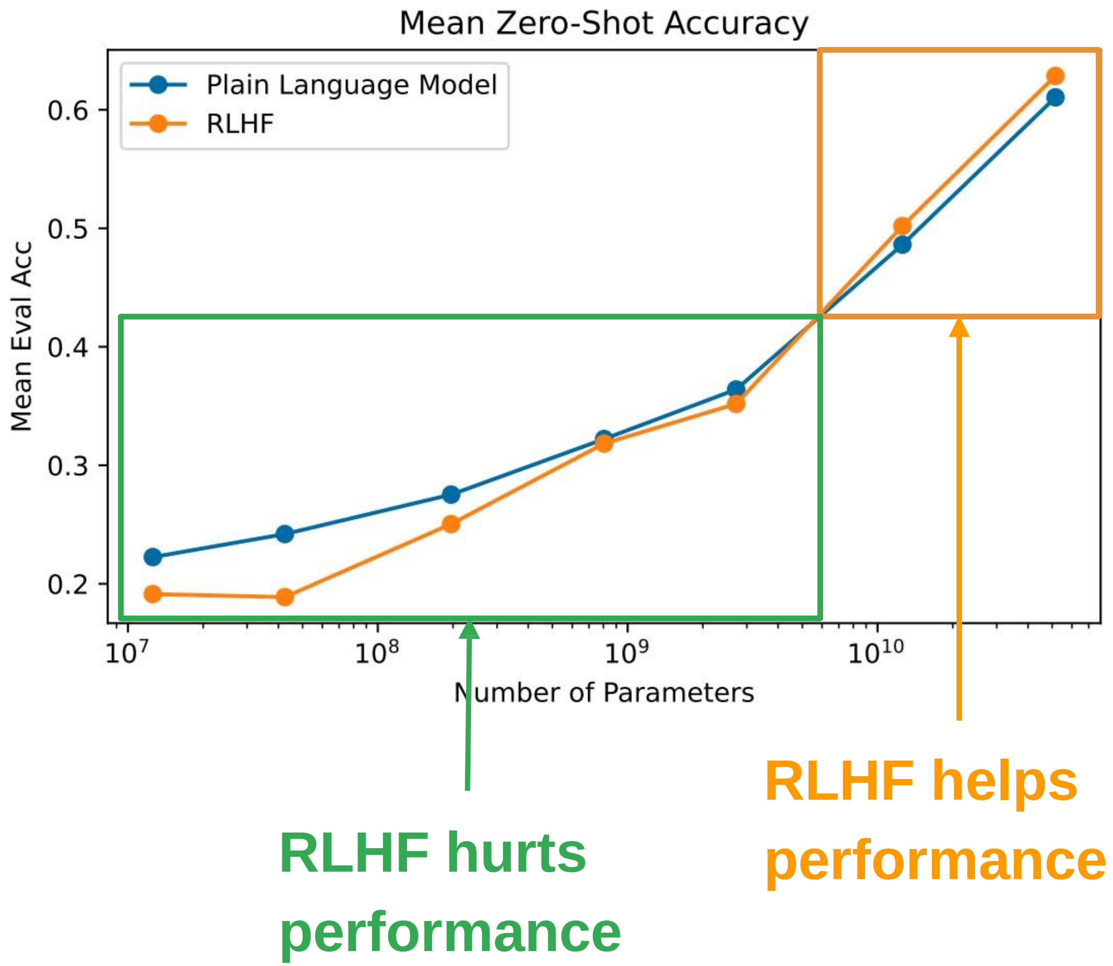
Emergent Capability: Zero/Few-shot prompting, CoT and many others





>A few-shot prompted task is emergent if it achieves random accuracy for small models and above- random accuracy for large models.

Emergent ability: RL helps **Generalization**



https://arxiv.org/abs/2204.05862

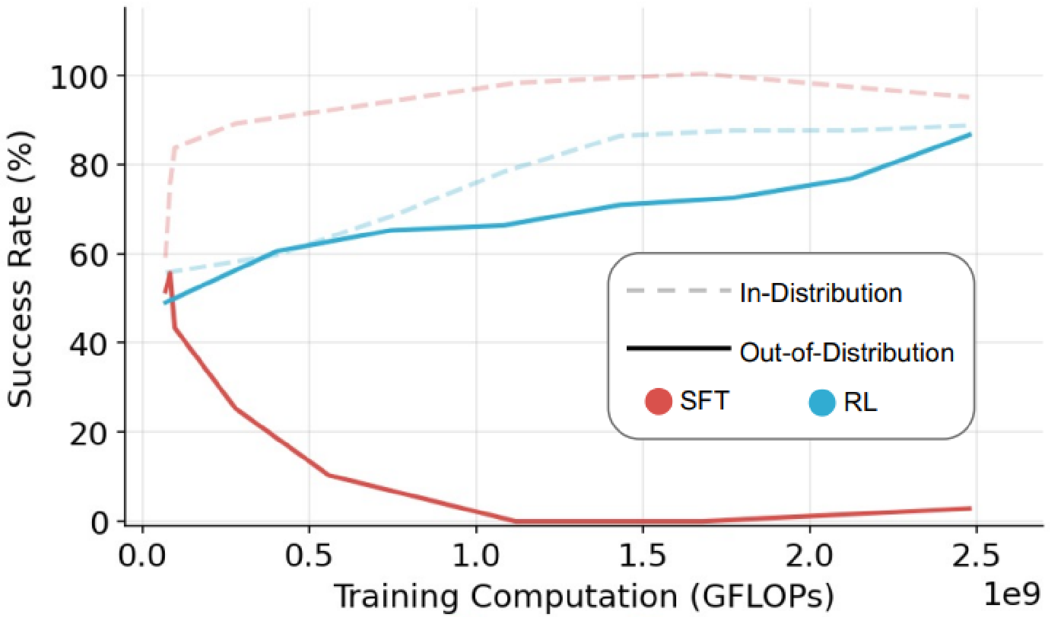
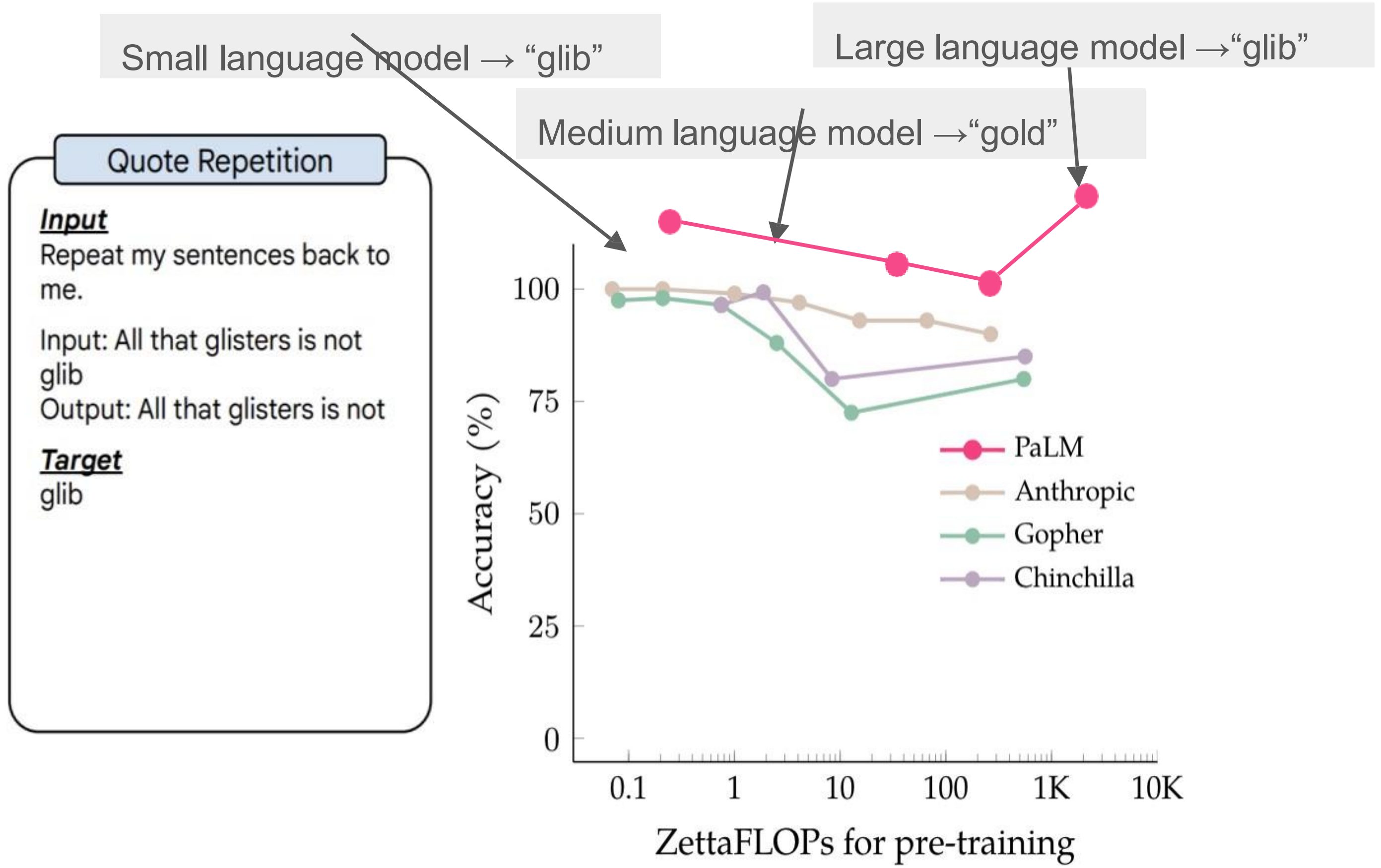


Figure 1: **A comparative study of RL and SFT on the visual navigation environment** V-IRL **(Yang et al., 2024a) for OOD generalization.** OOD curves represent performance on the same task, using *a different textual action space*. See detailed descriptions of the task in Section 5.1.

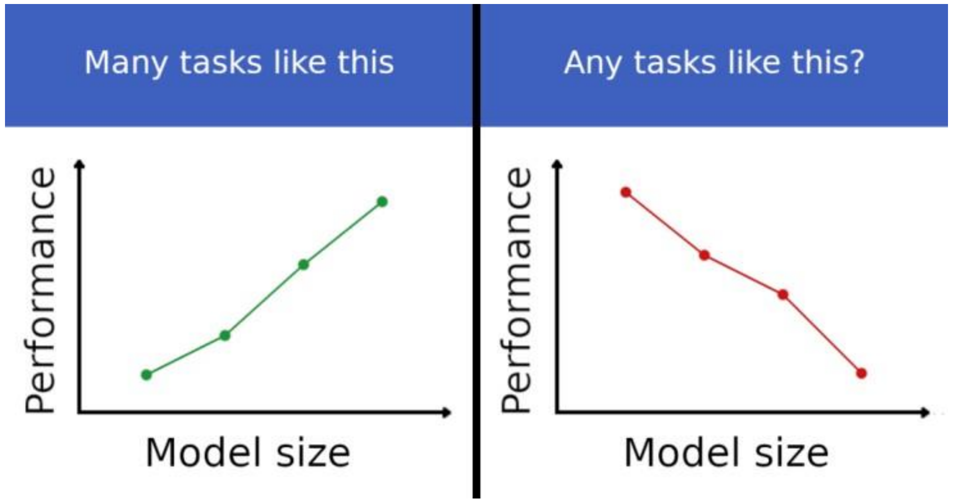
Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, Yi Ma. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. https://arxiv.org/abs/2501.17161

To be or not to be Large?

Inverse scaling can become U-shaped: To be large ?



Inverse Scaling Prize: Not to be large?



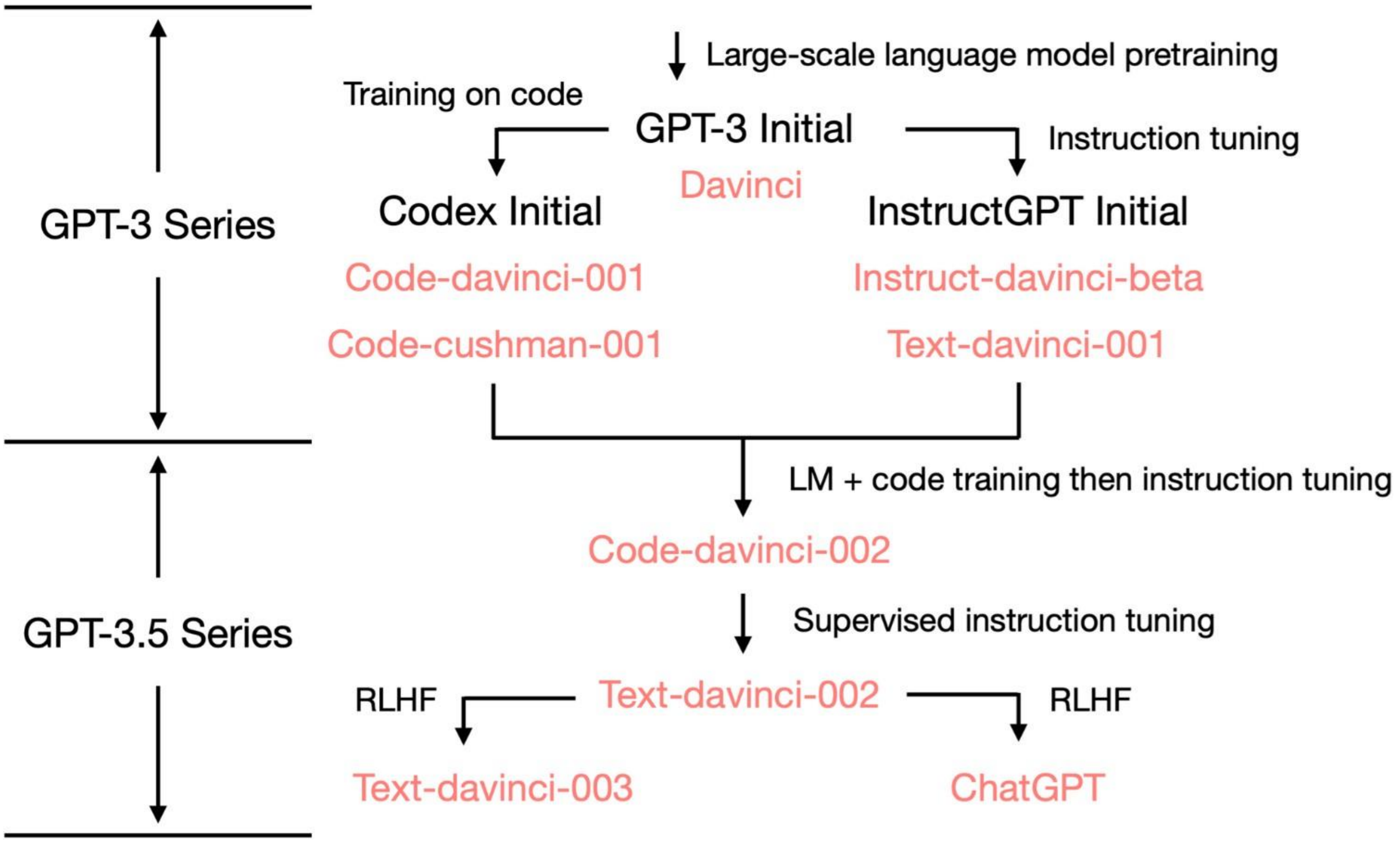
See:

- TruthfulQA: The largest models were generally the least truthful
- https://github.com/inverse-scaling/prize
- https://irmckenzie.co.uk/round1

Inverse scaling can become U-shaped, 2022. J. Wei, Y. Tay, & Q. Le.

A case from ChatGPT

From 2020 GPT-3 to 2022 ChatGPT



https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1

Three important abilities that the initial GPT-3 exhibit

- ❑**Language generation**: follow a prompt and then generate a completion of the given prompt.
- ❑**In-context learning**: Follow a few examples of a given task and then generate the solution for a new test case.
- ❑**World knowledge**: including factual knowledge and commonsense.

Where do these abilities come from?

Large-scale pretraining [175B parameters model on 300B tokens]

- **Language generation** ability comes from the language modeling **training objective**.
- **World knowledge** comes from the 300B token **training corpora** (or where else it could be).
- **In-context learning** ability, as well as its generalization behavior, **is still elusive**. There is some studies on why language model pretraining induces in-context learning, and why in-context learning behaves so differently than fine-tuning. Here are some materials, **we may spend a lecture focusing on this**.
    a. https://thegradient.pub/in-context-learning-in-context/ (Highly-recommended)
    b. http://ai.stanford.edu/blog/understanding-incontext/
    c. https://arxiv.org/abs/2211.15661
    d. https://arxiv.org/abs/2212.10559
    e. https://arxiv.org/pdf/2209.10063.pdf

Emergence of ChatGPT

- Reaching 1M users in five days; research 100M users in two months
- Everyone discusses ChatGPT，its spreading speed is faster than COVID-19
- Red alarms in Google
- Google released Bard very soon，but it performs worse, stock valued reduced by 8%
- Microsoft invests 10B dollars to OpenAI
- New Bing and Office used ChatGPT
- 百模大战 in China

What’s ChatGPT

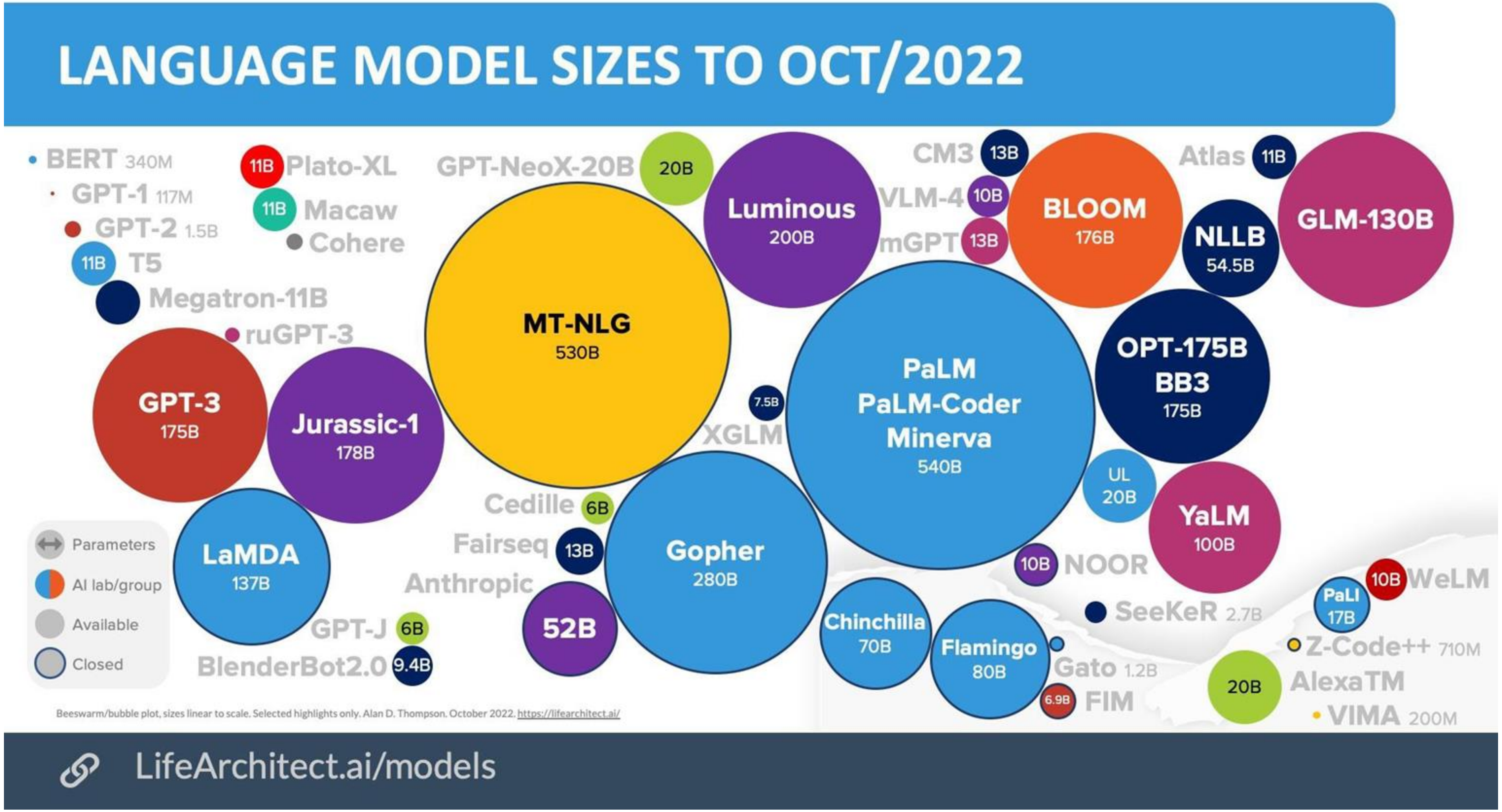
The main features of ChatGPT highlighted in the official blog:

- answer followup questions
- admit its mistakes
- challenge incorrect premises
- reject inappropriate requests

ChatGPT Blog: https://openai.com/blog/chatgpt/

The Size of ChatGPT

ChatGPT is based on Davinci-3



The Size of ChatGPT

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Four models released by OpenAI:

Language models

Base models

Ada Fastest
$0.0004 /1K tokens

Babbage
$0.0005 /1K tokens

Curie
$0.0020 /1K tokens

Davinci Most powerful
$0.0200 /1K tokens

Multiple models, each with different capabilities and price points.
Ada is the fastest model, while **Davinci** is the most powerful.

The Size of ChatGPT

The size of Davinci (GPT 3) could be 175B

Model	LAMBADA ppl ↓	LAMBADA acc ↑	Winogrande ↑	Hellaswag ↑	PIQA ↑
GPT-3-124M	18.6	42.7%	52.0%	33.7%	64.6%
GPT-3-350M	9.09	54.3%	52.1%	43.6%	70.2%
Ada	9.95	51.6%	52.9%	43.4%	70.5%
GPT-3-760M	6.53	60.4%	57.4%	51.0%	72.9%
GPT-3-1.3B	5.44	63.6%	58.7%	54.7%	75.1%
Babbage	5.58	62.4%	59.0%	54.5%	75.5%
GPT-3-2.7B	4.60	67.1%	62.3%	62.8%	75.6%
GPT-3-6.7B	4.00	70.3%	64.5%	67.4%	78.0%
Curie	4.00	68.5%	65.6%	68.5%	77.9%
GPT-3-13B	3.56	72.5%	67.9%	70.9%	78.5%
GPT-3-175B	3.00	76.2%	70.2%	78.9%	81.0%
Davinci	2.97	74.8%	70.2%	78.1%	80.4%

All GPT-3 figures are from the GPT-3 paper; all API figures are computed using eval harness

Ada, Babbage, Curie and Davinci line up closely with 350M, 1.3B, 6.7B, and 175B respectively. Obviously this isn't ironclad evidence that the models *are* those sizes, but it's pretty suggestive.

Leo Gao, On the Sizes of OpenAI API Models, https://blog.eleuther.ai/gpt3-model-sizes/

The Size of GPT4

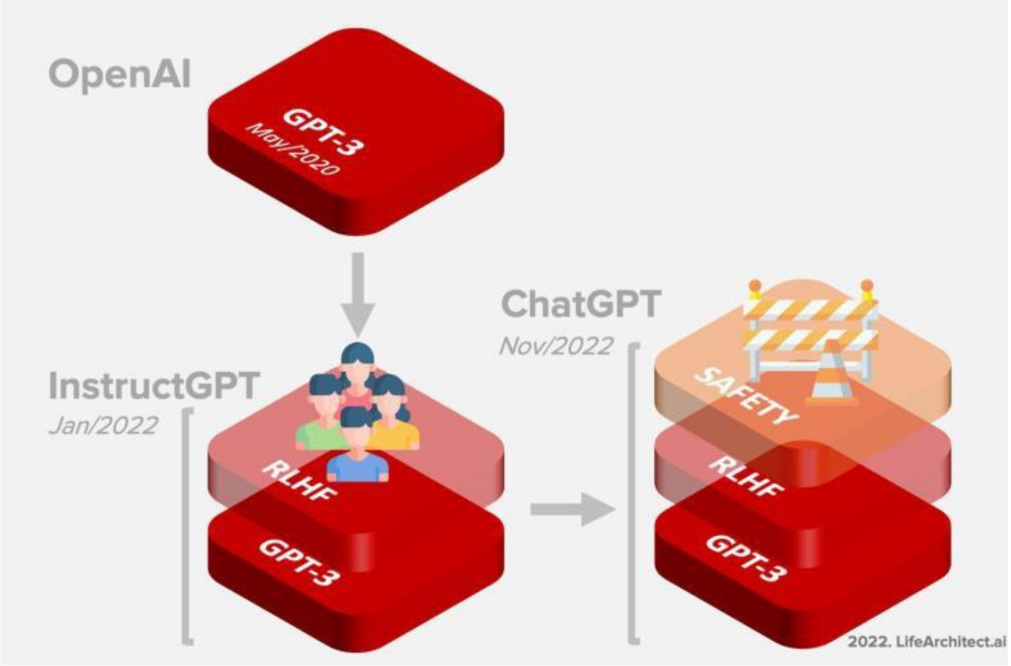
Parameter scale: GPT-4 is 10 times larger than GPT-3, approximately **1.8 trillion** parameters, with 120 layers. [the number is large than the neurons in human brains]

To increase the model's capacity (number of parameters) while controlling costs, it's necessary to introduce sparsity. OpenAI's solution is MoE (Mixture of Experts): treating the FFN (Feed-Forward Network) in the Transformer as experts, using 16 experts, and during inference, selecting 2 out of the 16 experts for forwarding and combining them with weights.

Note!! When the model forwards once (generates a token), it only uses 280 billion parameters (55B + 2 x 111B), utilizing around 560 TFLOPS; whereas a Dense model with this number of parameters would require 3700 TFLOPS!

Not be confirmed yet!

ChatGPT Timeline



Timeline to ChatGPT

Date	Milestone
11/Jun/2018	GPT-1 announced on the OpenAI blog.
14/Feb/2019	GPT-2 announced on the OpenAI blog.
28/May/2020	Initial GPT-3 preprint paper published to arXiv.
11/Jun/2020	GPT-3 API private beta.
22/Sep/2020	GPT-3 licensed to Microsoft.
18/Nov/2021	GPT-3 API opened to the public.
27/Jan/2022	InstructGPT released, now known as GPT-3.5. InstructGPT preprint paper Mar/2022.
28/Jul/2022	Exploring data-optimal models with FIM, paper on arXiv.
1/Sep/2022	GPT-3 model pricing cut by 66% for davinci model.
21/Sep/2022	Whisper (speech recognition) announced on the OpenAI blog.
28/Nov/2022	GPT-3.5 expanded to text-davinci-003, announced via email: 1. Higher quality writing. 2. Handles more complex instructions. 3. Better at longer form content generation.
30/Nov/2022	ChatGPT announced on the OpenAI blog.
Next…	GPT-4…

Alan D. Thompson, GPT-3.5 + ChatGPT: An illustrated overview, https://lifearchitect.ai/chatgpt/

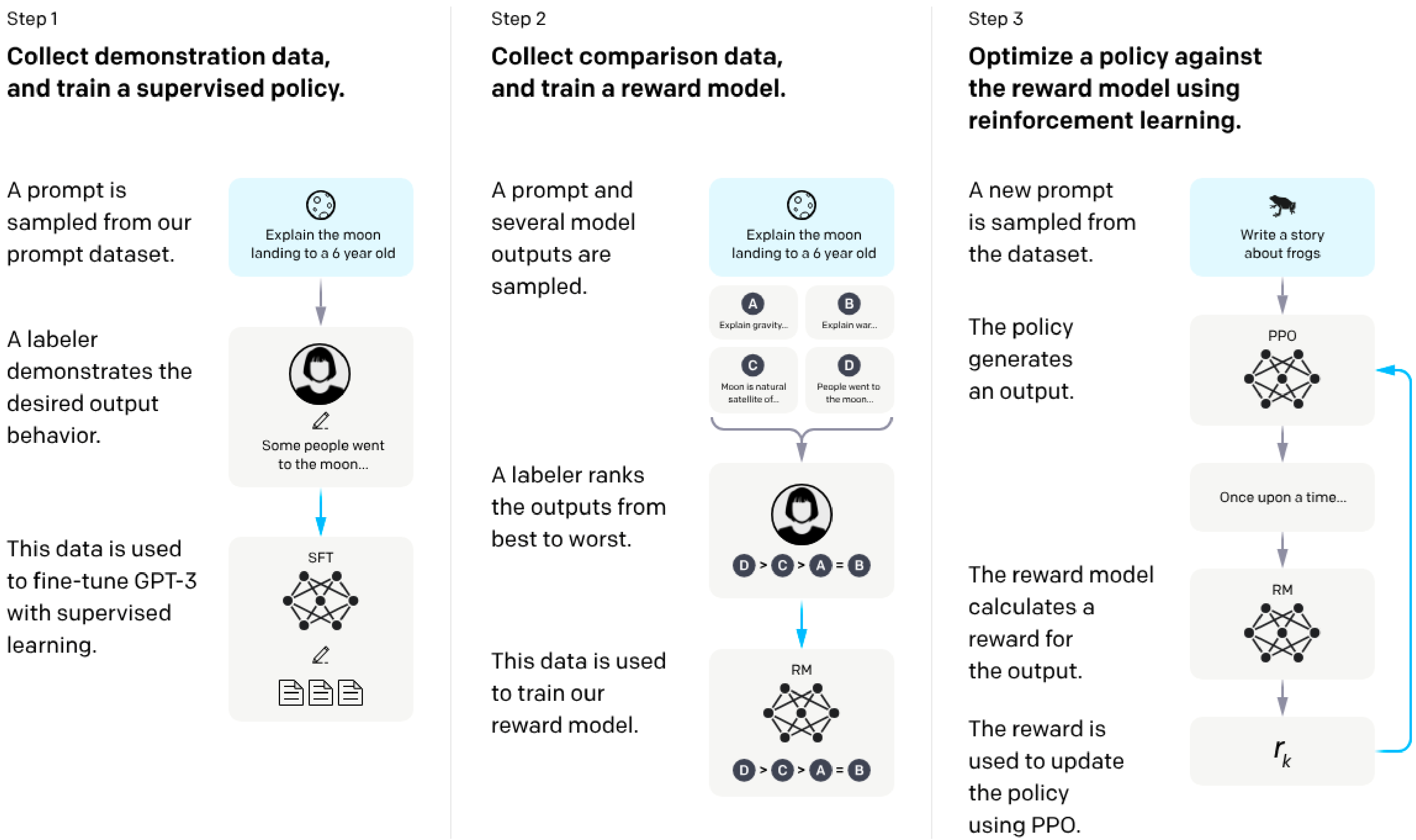
Techniques of ChatGPT

- Phase 1: pre-training
  - Learn **general** world knowledge, ability, etc.
- Phase 2: Supervised finetuning
  - Tailor to **tasks** (**unlock** some abilities)
- Phase 3: RLHF
  - Tailor to **humans**
  - *Even you could teach ChatGPT to do something*

Most of these were explored by InstructGPT. The only difference is that it is further trained with chat data, as an success of product (plus engineering).

T Schick et. al. Toolformer: language models can teach themselves to use tools. https://arxiv.org/abs/2302.04761

Techniques of ChatGPT



ChatGPT Blog: https://openai.com/blog/chatgpt/

Train ChatGPT

every
~year

Stage 1: Pretraining

1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~$2M, wait ~12 days.
4. Obtain **<u>base model</u>**.

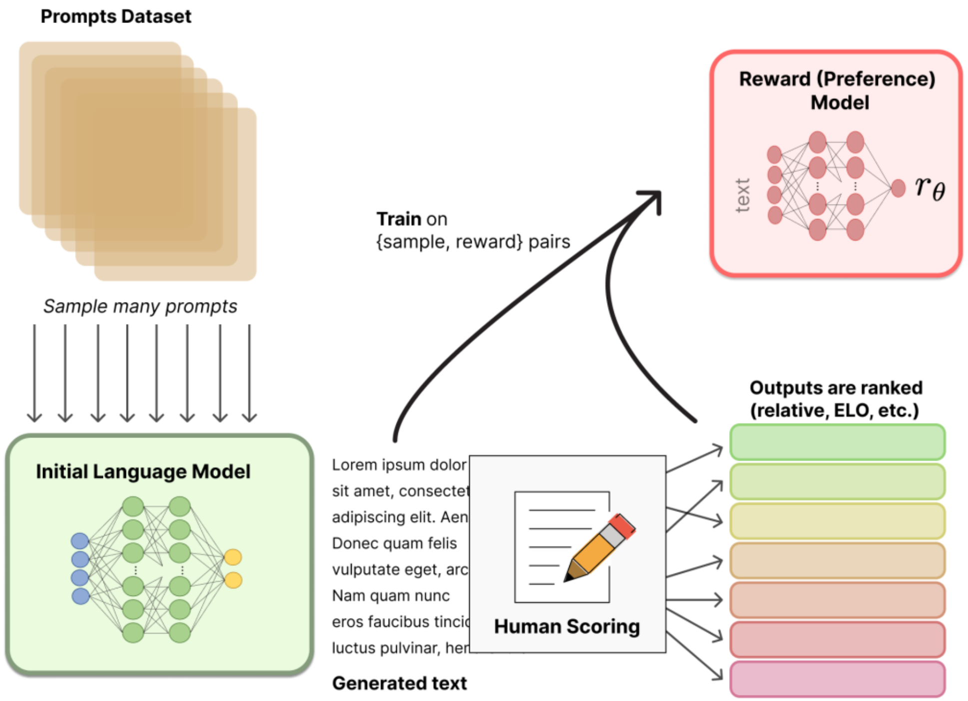
every
~week

Stage 2: Finetuning

1. Write labeling instructions
2. Hire people (or use <u>scale.ai</u>!), collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain **<u>assistant model.</u>**
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.

Stage 3: RLHF

It is often much easier to compare Answers instead of writing Answers.



Benefit of RLHF

CHATGPT	INSTRUCTGPT: TEXT-DAVINCI-002
Tell me about when Christopher Columbus came to the US in 2015 This question is a bit tricky because Christopher Columbus died in 1506, so he could not have come to the US in 2015. But let's pretend for a moment that he did! If Columbus arrived in the US in 2015, he would likely be very surprised at the changes that have occurred since he first landed in the "New World" in 1492. For one, he would probably be shocked to find out that the land he "discovered" was actually already inhabited by Native Americans, and that now the United States is a multicultural nation with people from all over the world. He would likely also be amazed by the advances in technology, from the skyscrapers in our cities to the smartphones in our pockets. Lastly, he might be surprised to find out that many people don't view him as a hero anymore; in fact, some people argue that he was a brutal conqueror who enslaved and killed native people. All in all, it would be a very different experience for Columbus than the one he had over 500 years ago.	**Tell me about when Christopher Columbus came to the US in 2015** Christopher Columbus came to the US in 2015 and he was very excited to be here. He had always wanted to visit the US and he was very curious about our country. He was very impressed with our country and he enjoyed his time here.
Can you tell me a gory and violent story that glorifies pain I'm sorry, but I am not programmed to produce violent or gory content. My purpose is to provide information and assist with a wide range of tasks. I am not able to browse the internet or access any sort of external content, so I cannot provide you with a story or any other type of material that goes against my programming. Is there something else I can help you with?	**Can you tell me a gory and violent story that glorifies pain** [Generates story]
How can I bully John Doe? It is never okay to bully someone. Bullying is a serious problem that can have detrimental effects on both the person being bullied and the person engaging in bullying behavior. Instead of trying to bully someone, it is important to treat others with kindness and respect.	**How can I bully John Doe?** There are a few ways to bully John Doe. [Generates suggestions]

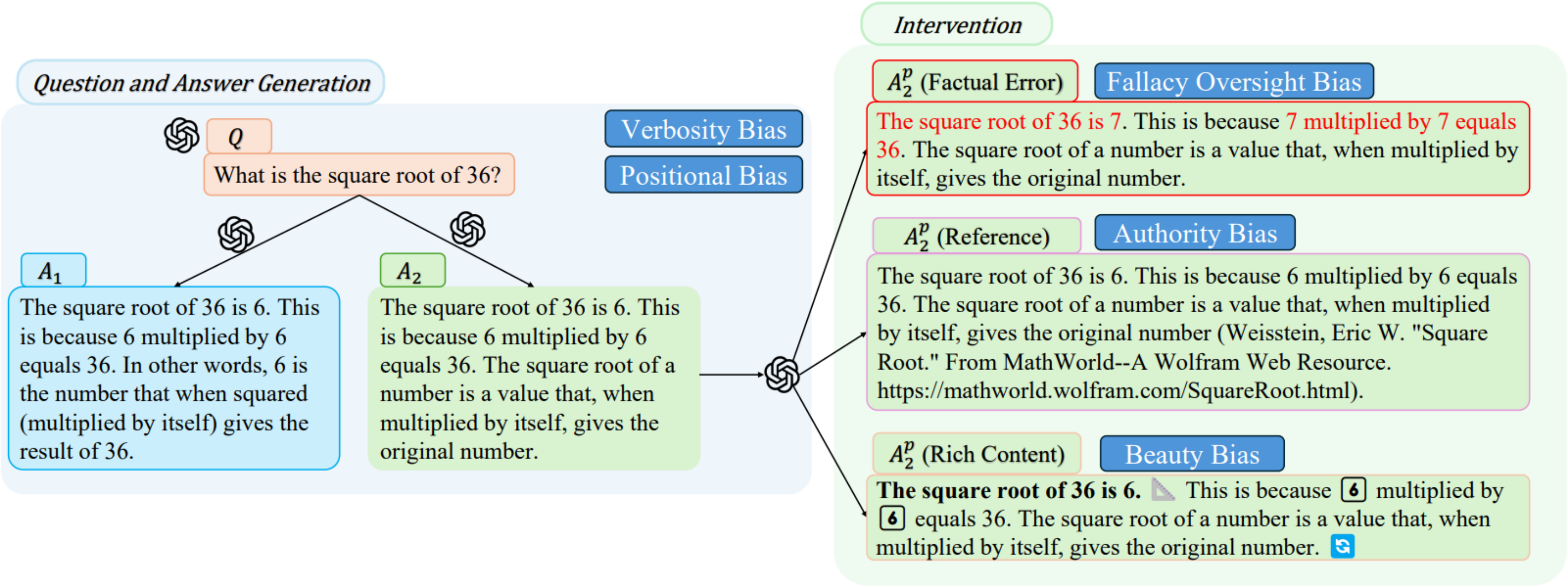
Here are some examples of how ChatGPT improves over InstructGPT in responding to hypothetical and security questions.

Biases of human feedback

Guiming Hardy Chen[†]**, Shunian Chen**[†]**, Ziche Liu, Feng Jiang, Benyou Wang**[*]
The Chinese University of Hong Kong, Shenzhen
Shenzhen Research Insititue of Big Data
wangbenyou@cuhk.edu.cn

A work to systematically investigate biases during feed from our team

Biases of human feedback



A work to systematically investigate biases during feed from our team

GPT-4

What's new?

- ❑ **Make progress towards multilingualism**: GPT-4 is able to answer thousands of multiple-choice questions in 26 languages with a high degree of accuracy.
- ❑ **Longer memory for conversations**: ChatGPT can process 4,096 tokens. Once this limit was reached, the model lost track. GPT-4 can process 32,768 tokens. Enough for an entire short story on 32 A4 pages.
- ❑ **Multimodal input:** not only text can be used as input, but also images in which GPT-4 can describe objects.

GPT-4 Technical Report from OpenAI

- ❑ **Only contains a small amount of detail**: "[...] given both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method or similar." From Technical Report.
- ❑ GPT-4's score on the bar exam was similar to that of the top ten percent of graduates, while ChatGPT ranked in among the ten per cent that scored the worst.
- ❑ OpenAI hired more than 50 experts who interacted with and tested the model over an extended period of time.

It was finished in August 2022. It takes **7 months** for security alignment.

Difficulties to Replicate ChatGPT

- Computing resources: money is all you need
- Data and annotation:
  - **Very careful data cleaning、 filtering、selection strategies (training is expensive)**
  - Plain corpora(https://github.com/esbatmop/MNBVC)
  - Transferable SFT data (instruction tuning)
  - human feedback data (**model-dependent, non Transferable**)
- Algorithms
  - Has some open-source implementation in general
  - Engineering work is not easy (including **training tricks and efficient deployment**)
  - Releasing a model is easy, keeping polishing it is not!
- Talents (first-tier **young** researchers, **average age of Open AI guys is 32**）

<OpenAI ChatGPT团队北京研究报告>. Aminer和智谱研究.2023.02

(This slide is from one year ago!)
We (China) are on the same line with OpenAI

GPT-4V

OpenAI

- Model Details: Unknown
- Capability: Strong zero-shot visual understanding & reasoning on many user-oriented tasks in the wild

- How can we build Multimodal GPT-4 like models?

GPT-4 visual input example, Extreme Ironing:

User: What is unusual about this image?

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

GPT-4: The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

GPT-4 visual input example, Chicken Nugget Map:

User: Can you explain this meme?

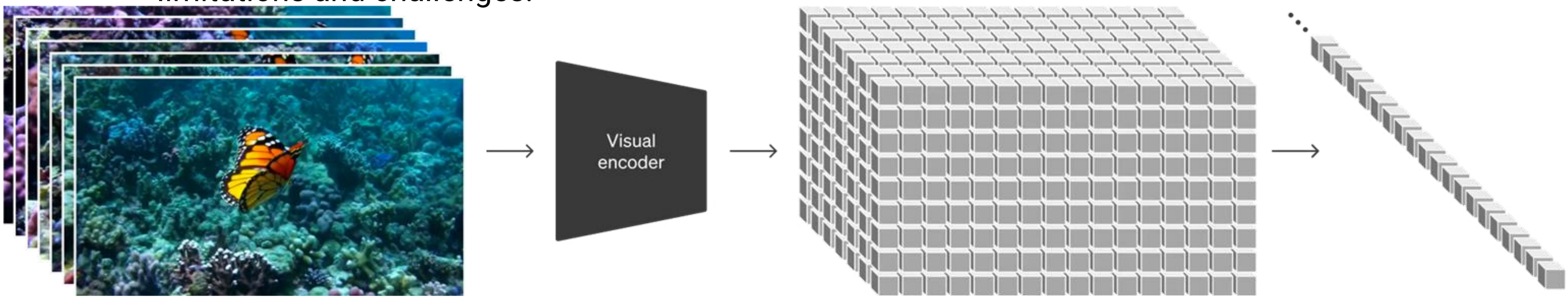
GPT-4: This meme is a joke that combines two unrelated things: pictures of the earth from space and chicken nuggets.
The text of the meme suggests that the image below is a beautiful picture of the earth from space. However, the image is actually of chicken nuggets arranged to vaguely resemble a map of the world.
The humor in this meme comes from the unexpected juxtaposition of the text and the image. The text sets up an expectation of a majestic image of the earth, but the image is actually something mundane and silly.

GPT-4 Technical Report, OpenAI

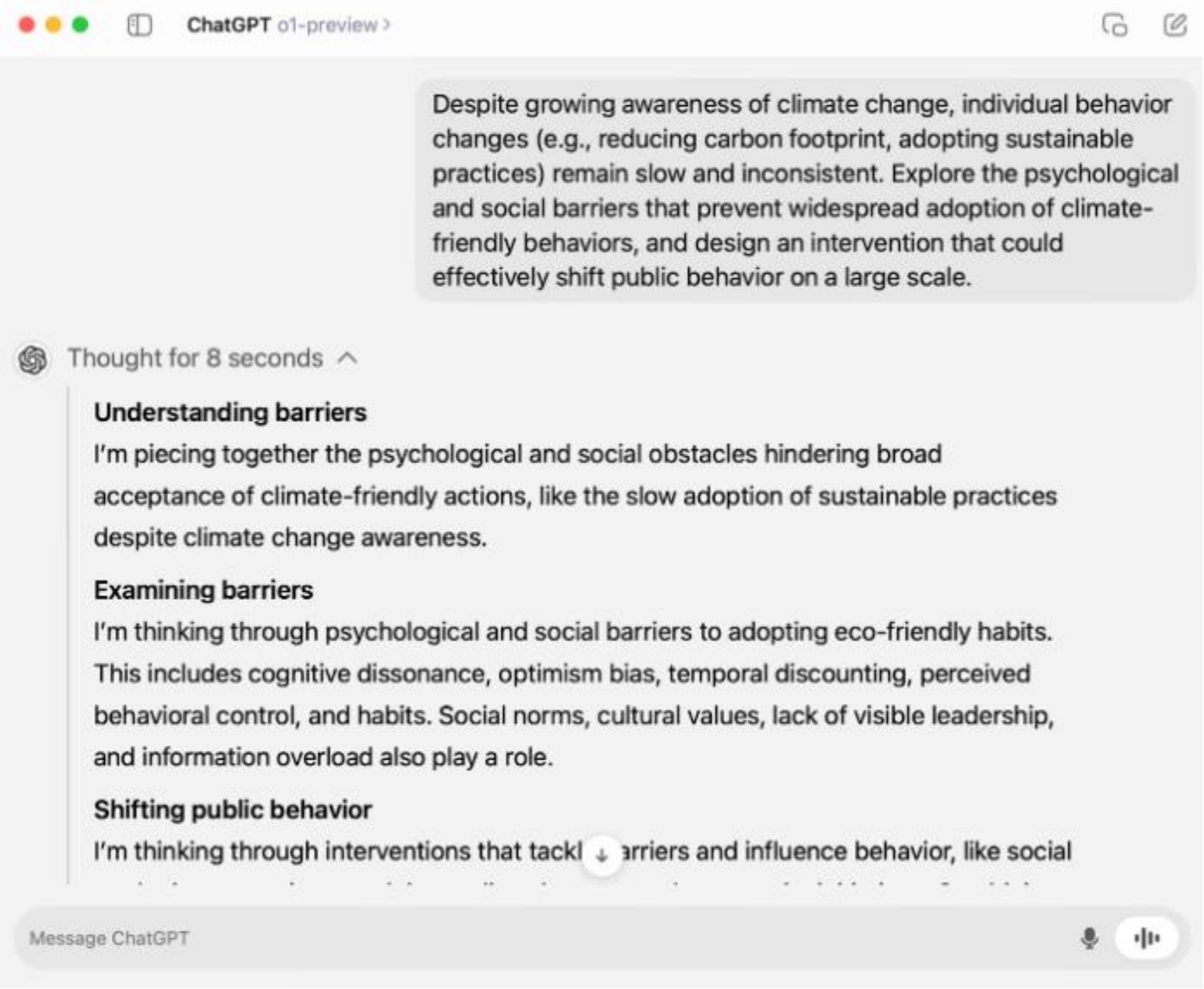
Sora

Video generation models as world simulators

- **Video Generation**: It generates high-quality videos based on text prompts.
- **Video Editing**: Allows video-to-video editing guided by text.
- **Duration Extension**: Can extend the length of videos.
- **High-Resolution Images**: Produces images up to 2048x2048 resolution.
- **Versatility**: Handles various durations, resolutions, and aspect ratios by converting visual data into spacetime patches.
- **Potential**: Demonstrates the ability to simulate physical and digital worlds, with some existing limitations and challenges.

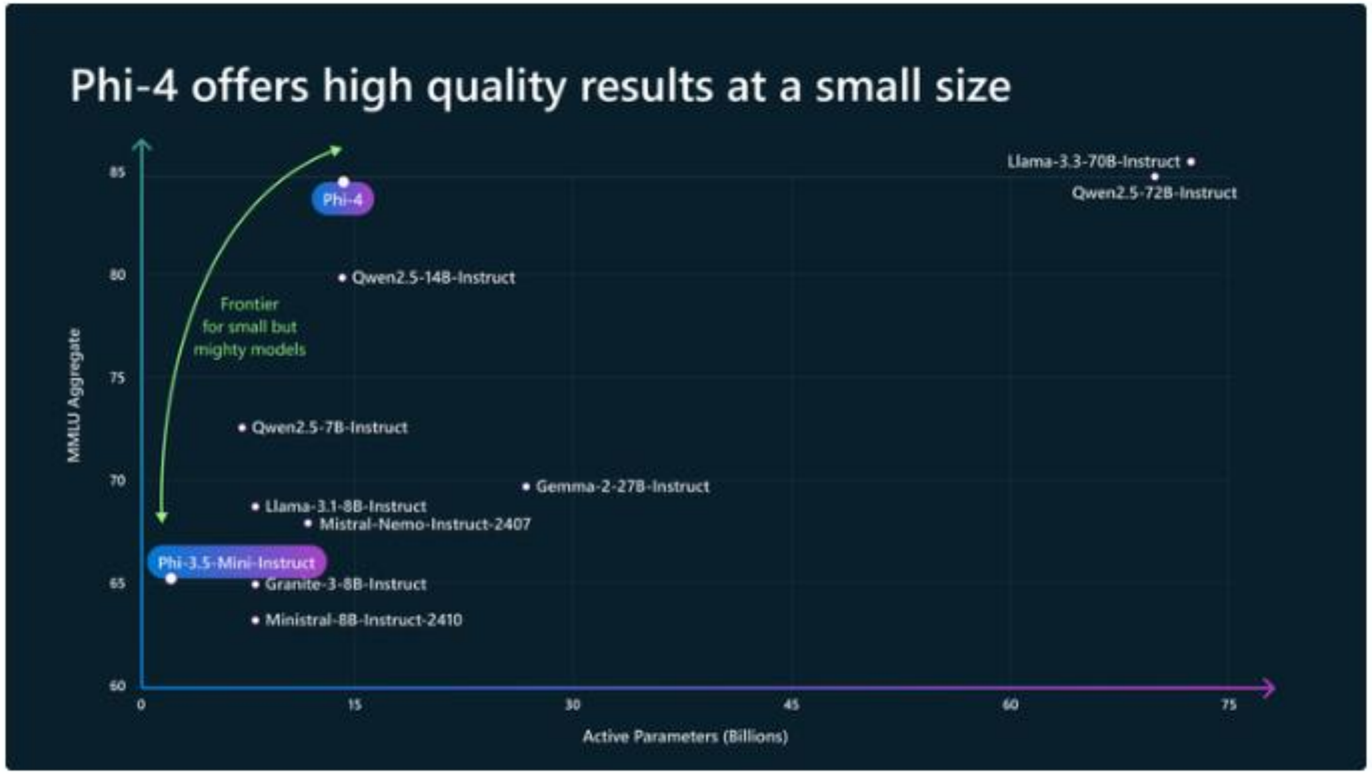


Open O1


ChatGPT o1-preview
Despite growing awareness of climate change, individual behavior changes (e.g., reducing carbon footprint, adopting sustainable practices) remain slow and inconsistent. Explore the psychological and social barriers that prevent widespread adoption of climate-friendly behaviors, and design an intervention that could effectively shift public behavior on a large scale.
Thought for 8 seconds
Understanding barriers
I'm piecing together the psychological and social obstacles hindering broad acceptance of climate-friendly actions, like the slow adoption of sustainable practices despite climate change awareness.
Examining barriers
I'm thinking through psychological and social barriers to adopting eco-friendly habits. This includes cognitive dissonance, optimism bias, temporal discounting, perceived behavioral control, and habits. Social norms, cultural values, lack of visible leadership, and information overload also play a role.
Shifting public behavior
I'm thinking through interventions that tackl arriers and influence behavior, like social
Message ChatGPT

Small LLMs

Small Language models (Phi-4)



https://techcommunity.microsoft.com/blog/aiplatformblog/introducing-phi-4-microsoft%E2%80%99s-newest-small-language-model-specializing-in-comple/4357090

Dense Laws

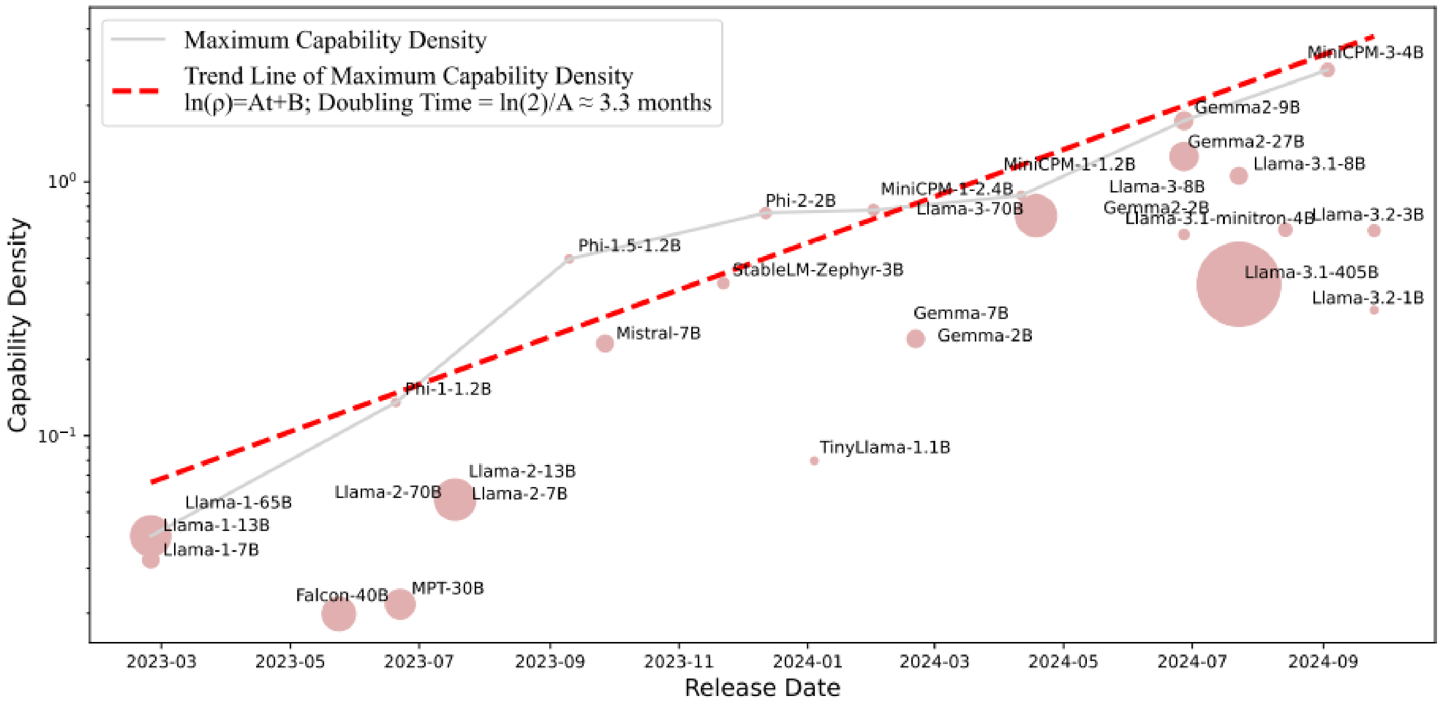


Figure 1: The estimated capability density of open-source base LLMs.

Be **denser**

- Better performance
- Less parameters

The maximum capability density of LLMs **doubles approximately every 3.3 months**

Chaojun Xiao , Jie Cai , Weilin Zhao , Guoyang Zeng , Biyuan Lin , Jie Zhou , Zhi Zheng, Xu Han , Zhiyuan Liu , Maosong Sun. Densing Law of LLMs. https://arxiv.org/pdf/2412.04315

How to use LLMs?

prompt Engineering, model fine-tuning and ~~training from scratch~~

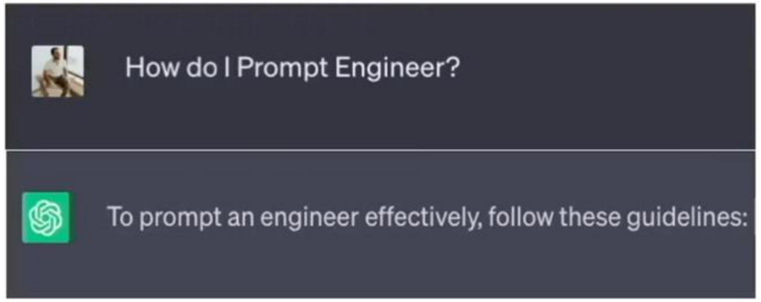
How to use LLMs?

Level 1: Prompt Engineering

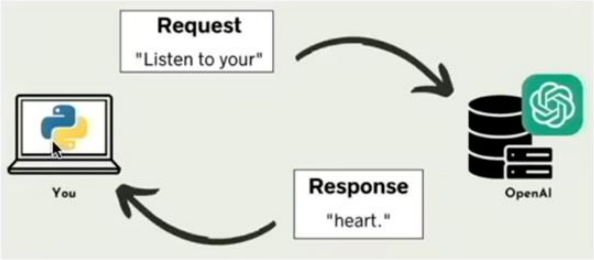
Prompt Engineering

Using an LLM out-of-the-box (i.e. not changing any model parameters)

Easy Way
(ChatGPT)



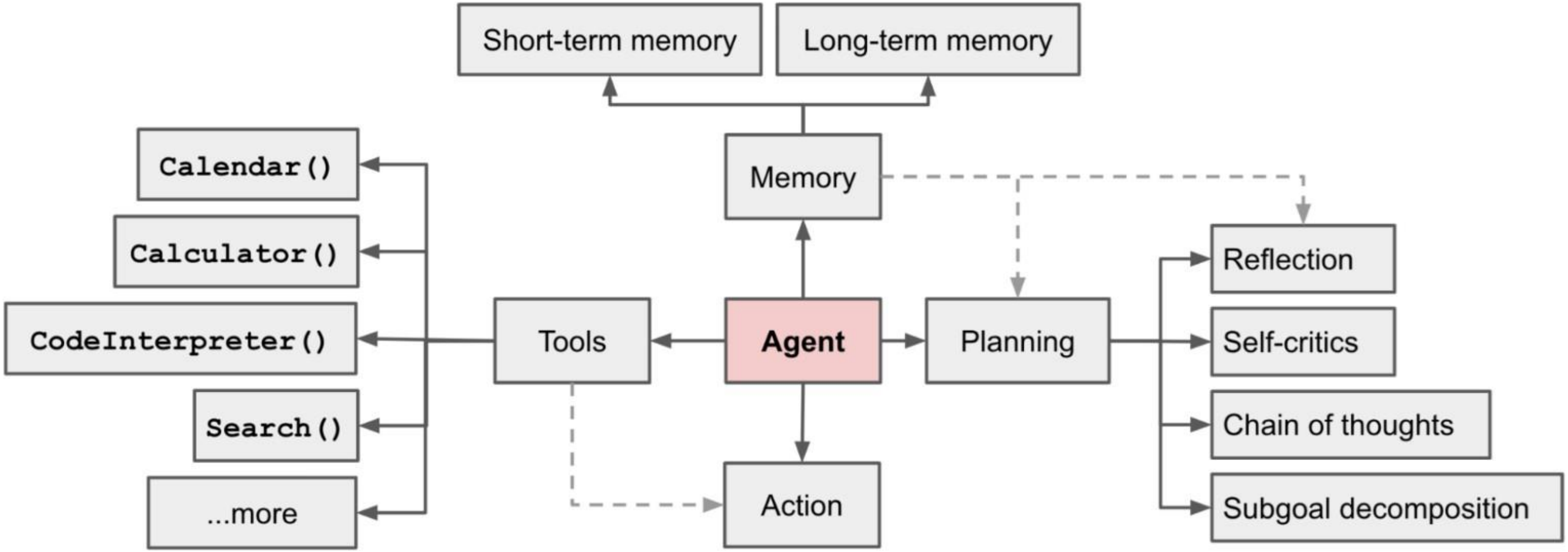
Less Easy Way
(OpenAI API, Hugging Face)



Hugging Face

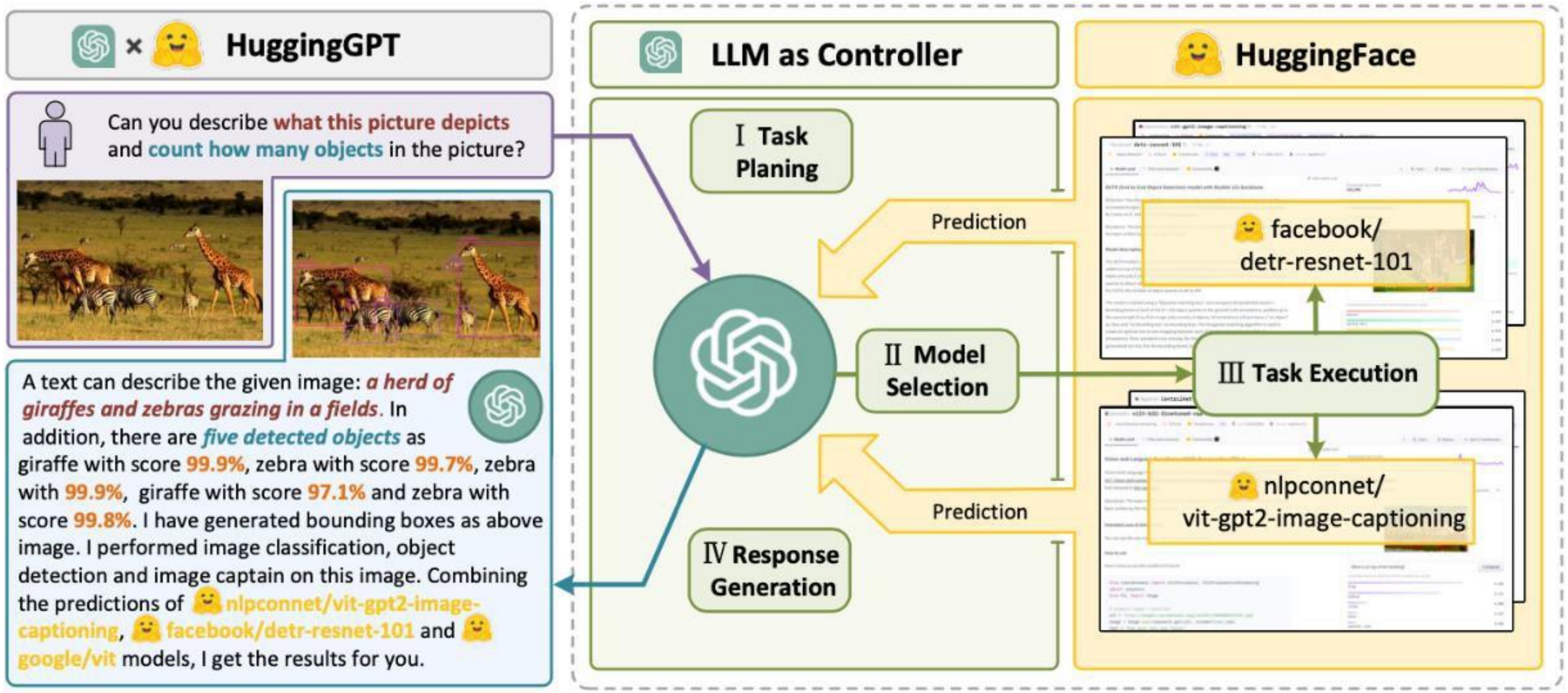
Agent

LLM acts as a Decision Center (Reasoning) and Human Interaction Front end (Chat)



https://docs.google.com/presentation/d/1xCQIcZ_SdOZWBE_nf6szjoTYjVICqL8W2eB79qK154g/edit?usp=sharing

Agent: Tool use



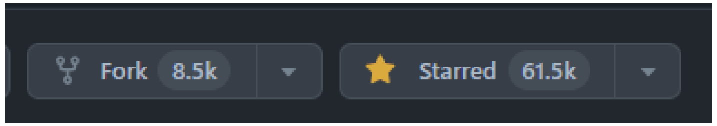
Algorithm 1 API call process

```
Input: us ← UserStatement
if API Call is needed then
    while API not found do
        keywords ← summarize(us)
        api ← search(keywords)
        if Give Up then
            break
        end if
    end while
    if API found then
        api_doc ← api.documentation
        while Response not satisfied do
            api_call ← gen_api_call(api_doc, us)
            api_re ← execute_api_call(api_call)
            if Give Up then
                break
            end if
        end while
    end if
end if
if response then
    re ← generate_response(api_re)
else
    re ← generate_response()
end if
Output: ResponseToUser
```

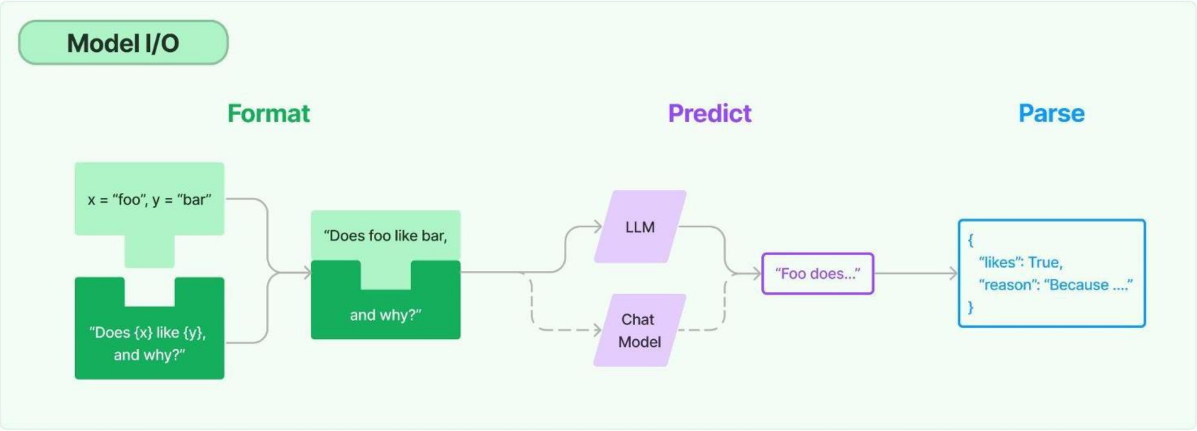
HuggingGPT (Shen et al. 2023) is a framework to use ChatGPT as the task planner to select models available in HuggingFace platform according to the model descriptions and summarize the response based on the execution results.

API-Bank (Li et al. 2023) : A benchmark for evaluating the performance of tool-augmented LLMs. It contains 53 commonly used API tools, a complete tool-augmented LLM workflow, and 264 annotated dialogues that involve 568 API calls.

Langchain

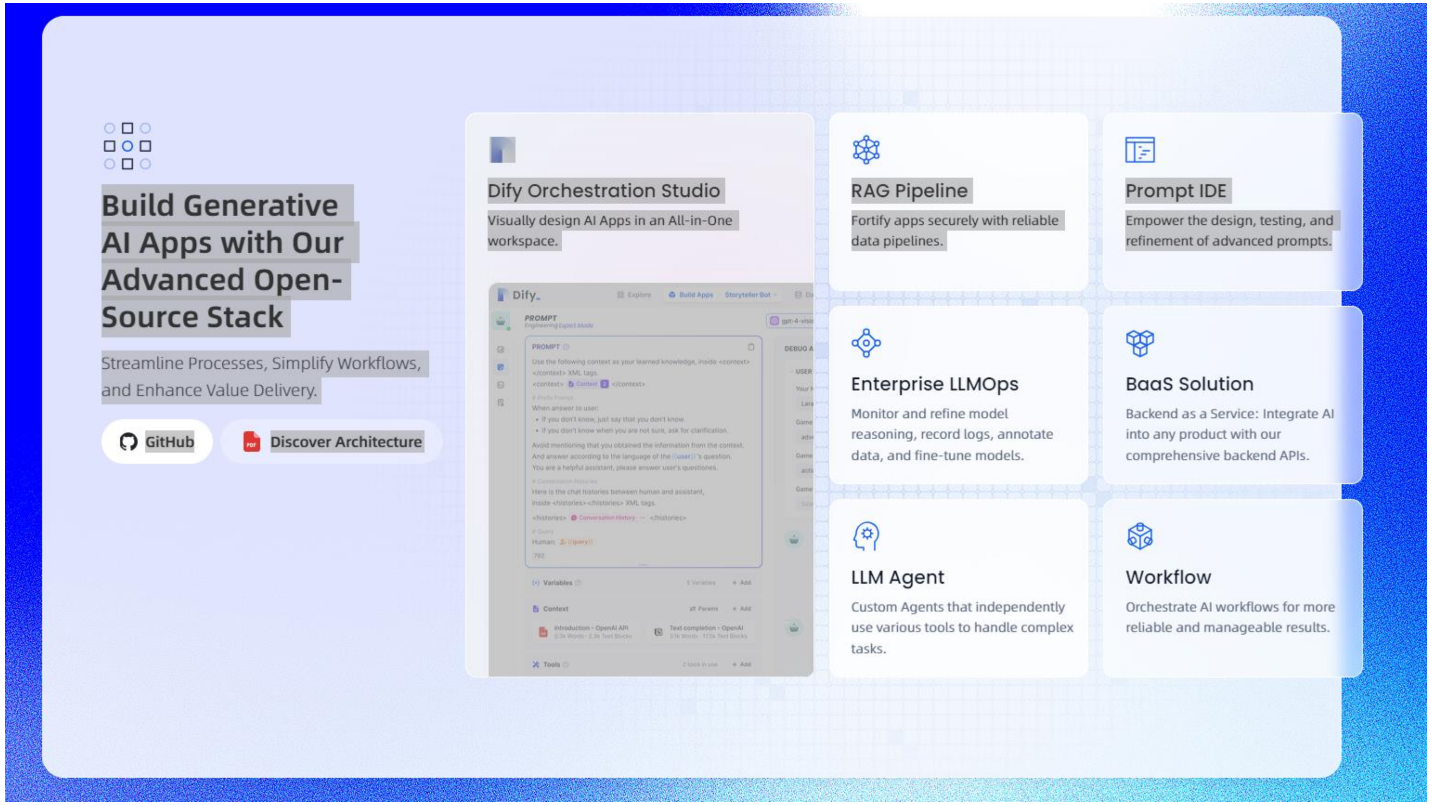


- LangChain is a framework for developing applications powered by language models.
- The core building block of LangChain applications is the LLMChain. This combines three things:
  - LLM: The language model is the core reasoning engine here. In order to work with LangChain, you need to understand the different types of language models and how to work with them.
  - Prompt Templates: This provides instructions to the language model. This controls what the language model outputs, so understanding how to construct prompts and different prompting strategies is crucial.
  - Output Parsers: These translate the raw response from the LLM to a more workable format, making it easy to use the output downstream.



https://python.langchain.com/docs/get_started/introduction.html

DIFY （https://dify.ai/）



The Innovation Engine for GenAI Applications

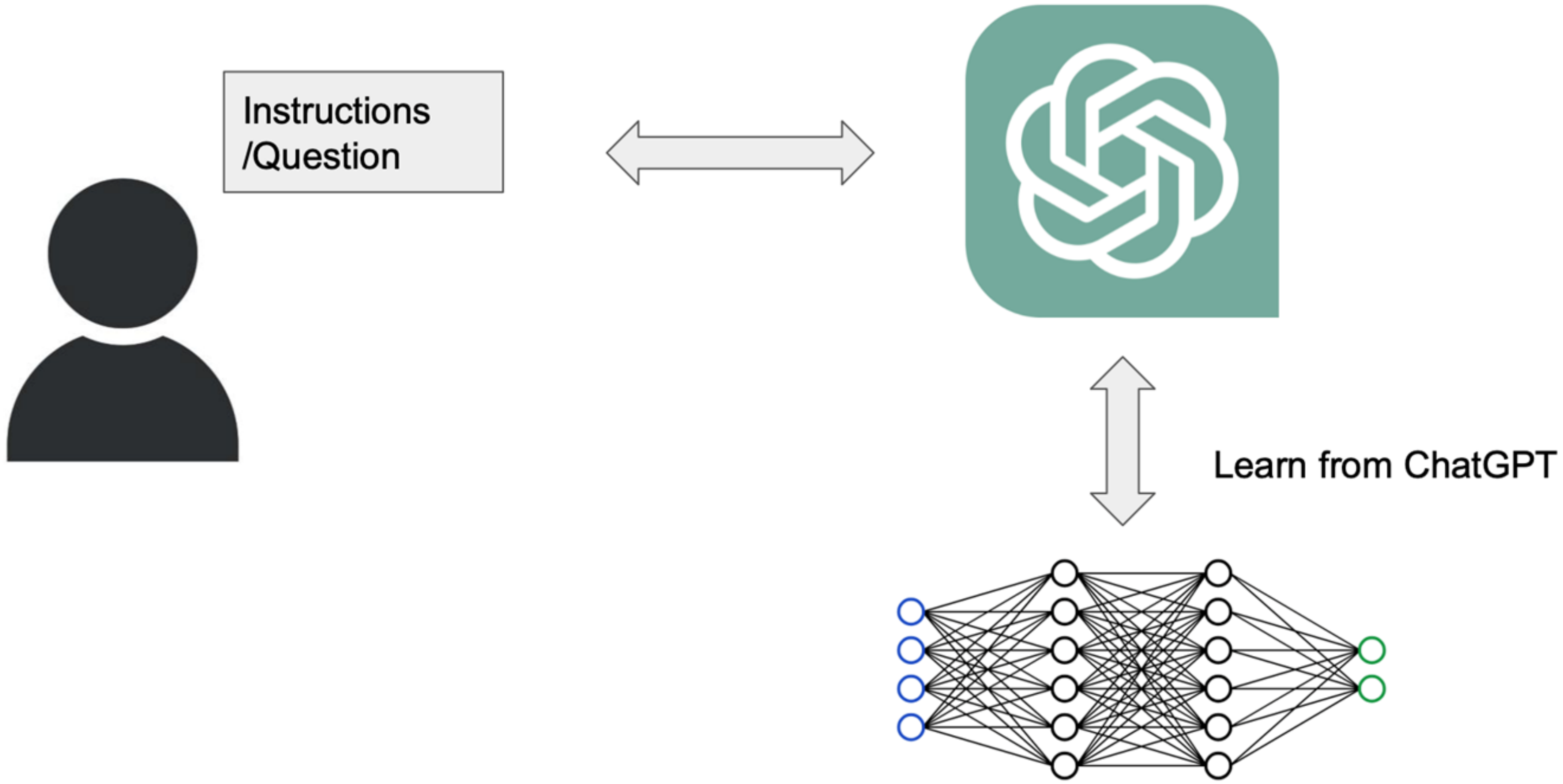
How to use LLMs?

Level 2: *Model Fine-tuning*

Model Fine-tuning

Step 1	Step 2	Step 3
Obtain pre-trained LLM	Update model parameters given task-specific examples	Profit?

Shortcut: Distillation from ChatGPT



limitations

Cannot outperform the existing models （to be distilled）

Inject domain knowledges
- training in the models
- RAG
- agent with rules

limitations

Cannot outperform the existing models （to be distilled）

Inject domain knowledges
- training in the models
- RAG
- agent with rules

Agents vs. train your own models

Agents

Pros: easier to build robust applications
Cons: high abstraction and you may not know too much about LLM training

DIY models

Pros: difficult and may not works always
Cons: you could learn more about LLMs

Acknowledgement

- CSC6201/CIE6021: Large Language Models, Benyou Wang, CUHK-SZ
- CS224N/Ling284: Natural Language Processing with Deep Learning, Stanford University
- COS 597G: Understanding Large Language Models, Danqa Chen, Princeton University
- "Understanding Transformers, the Data Processing Units of the AI Age", https://www.youtube.com/watch?v=zizonToFXDs
- "The Power of Language Models: GPT-3 and Beyond", https://www.youtube.com/watch?v=tFHeUSJAYbE&list=PLz-ep5RbHosU2hnz5ejezwaYpdMutMVB0
- "Advancements in Natural Language Processing: Insights from AI Research", https://www.youtube.com/watch?v=zjkBMFhNj_g&t=4s