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

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

Lecture 5-2: Pretraining and SFT

Spring 2025
Benyou Wang
School of Data Science

Before the lecture:

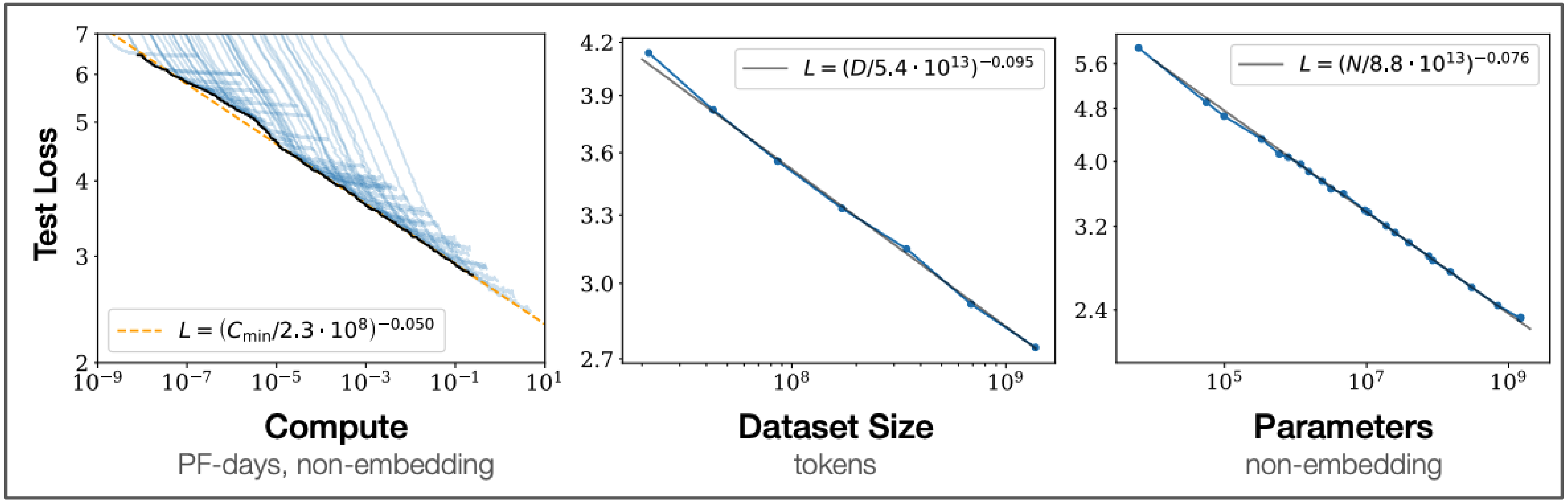
- GPT 4.5 is coming

Recap

- DL hypothesis:
  - Anything a human do in **0.1 seconds**, a big 10-layer neural network can do, too.
- Jason Wei' Rule of thumb
  - Language models can do (with decent accuracy) most things that an average human can do in **1 minute**.
- 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**.

Think about AGI?

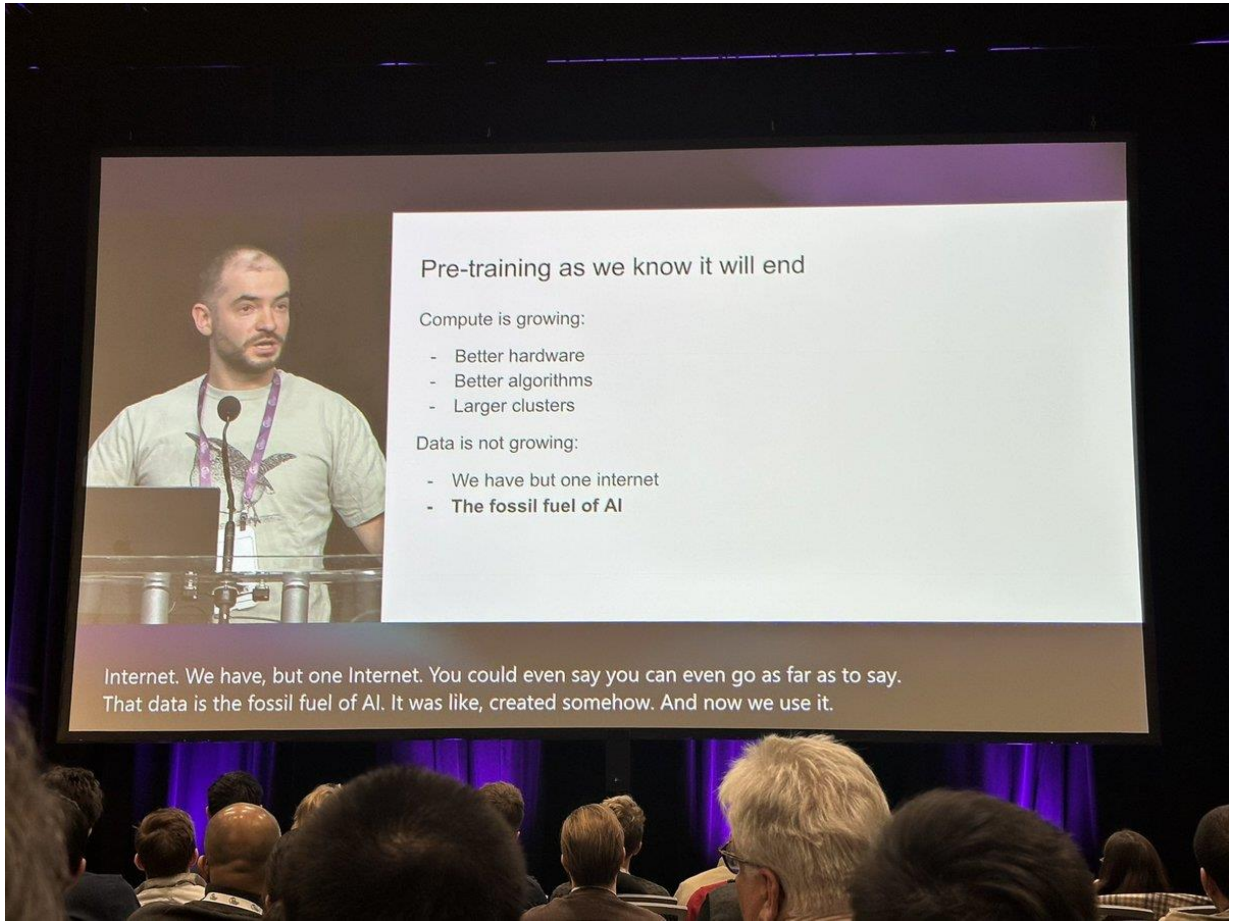
Benefits to be large： **Scaling Law?**



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

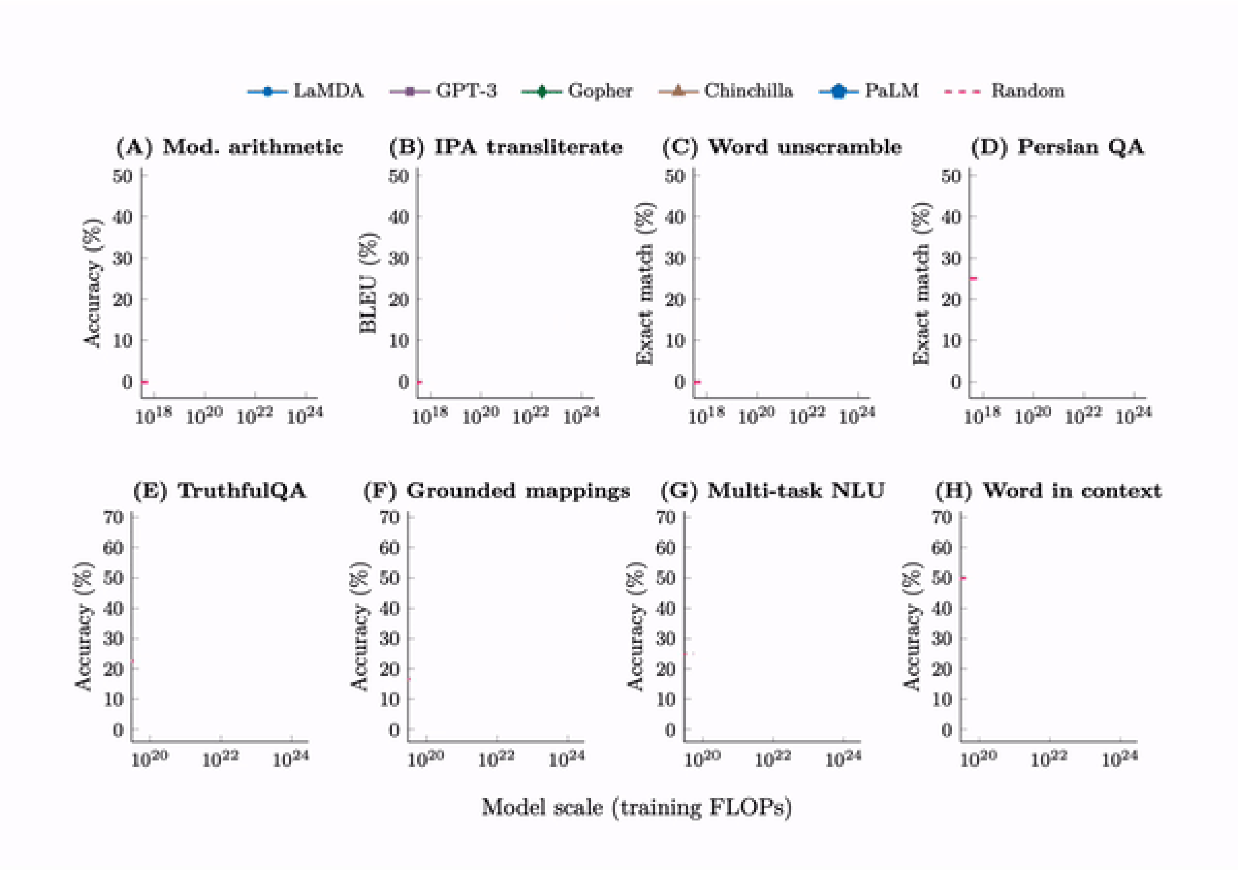
Emergent abilities of large language models (TMLR '22).
J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, & W. Fedus.

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



https://youtu.be/1yvBqasHLZs

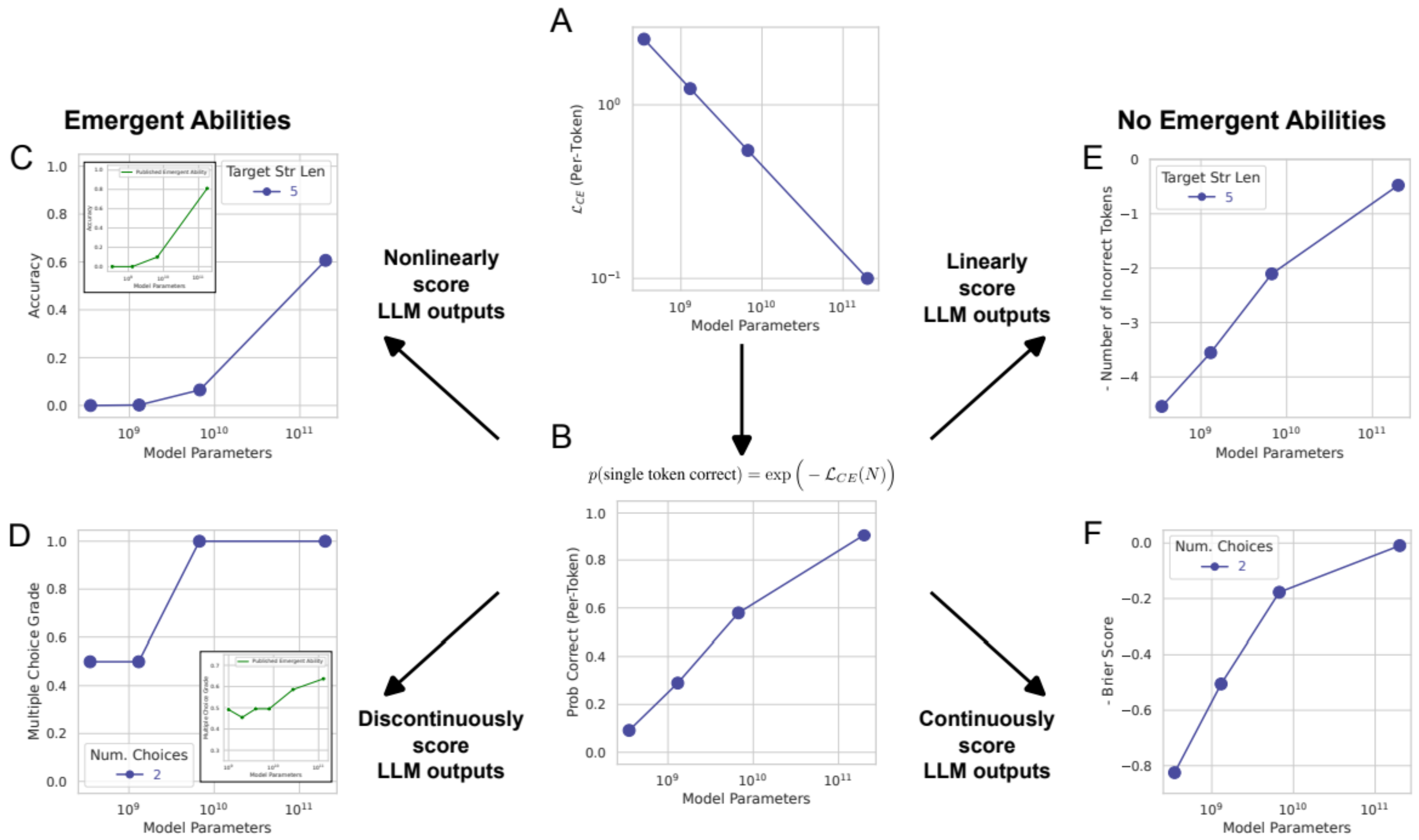
Benefits to be large: **Emergent ability?**



Some ability of LM (e.g. few-shot learning) is not present in smaller models but is present in larger 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

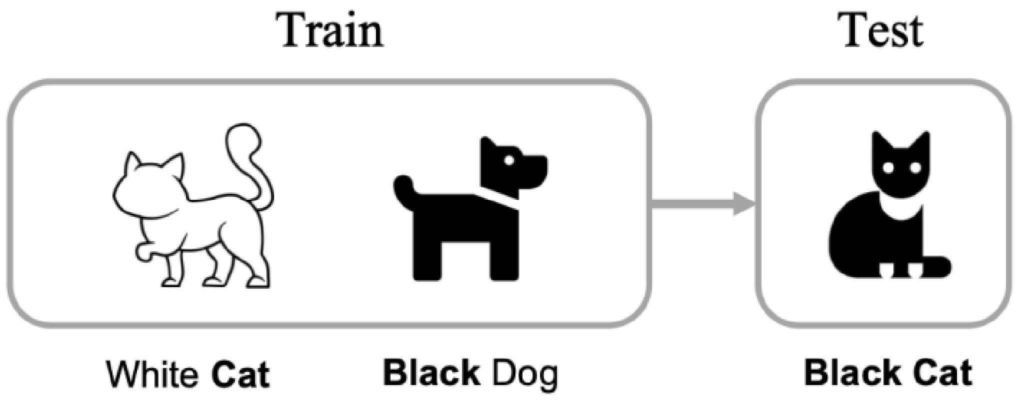


It seems that emergent ability of a model only occurs if the measure of per-token error rate of any model is scaled **non-linearly or discontinuously**.

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

My thought

- Larger capacity for **better generalization**
- Generalization might be attributed to **Combinational Generalization**, as it has seen all data during pretraining.



Enabling high-order **Combinational Generalization needs long thinking;**

The Future

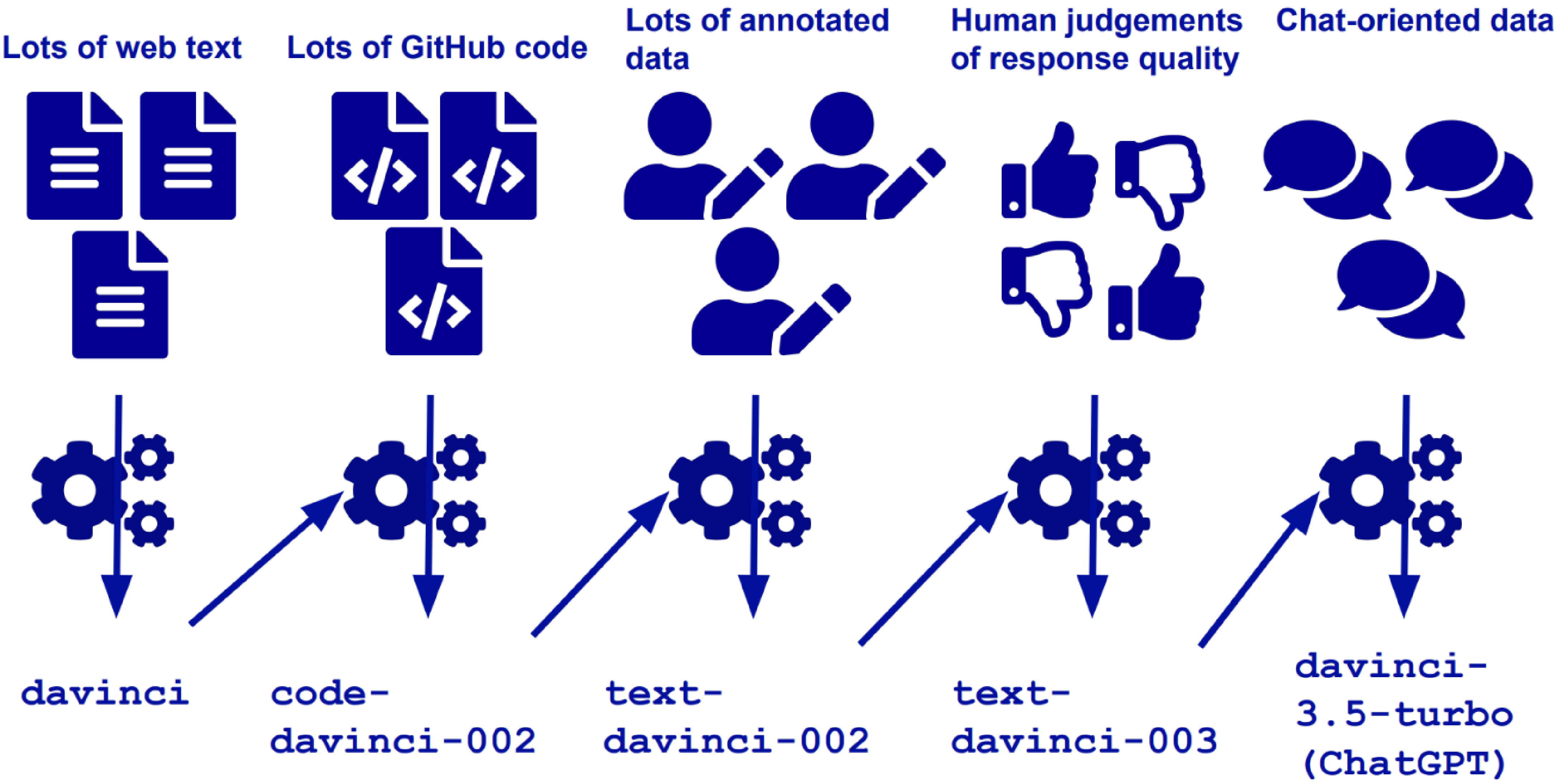
- **Data** is nearly over
  - “We only have on internet”, says Ilya Sutskever
- **Model** scales become saturated due to the hardware
  - A single GPU server (80*8) can only deploy a model up to 700B using INT8 quantization.

Scaling law -> **Densing** law!

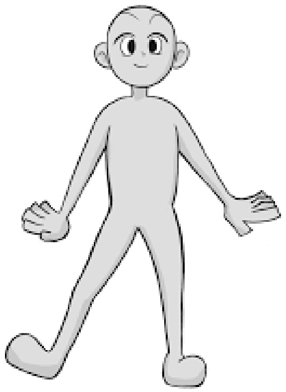
Chaojun Xiao , Jie Cai , Weilin Zhao , Guoyang Zeng , Biyuan Lin , Jie Zhou , Zhi Zheng, Xu Han , Zhiyuan Liu , Maosong Sun. Densing Law of LLMs. https://arxiv.org/pdf/2412.04315

Understanding of LLM Training

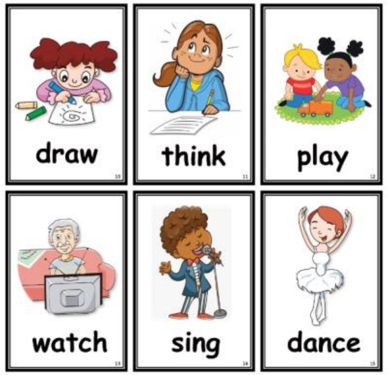
From Zero to ChatGPT

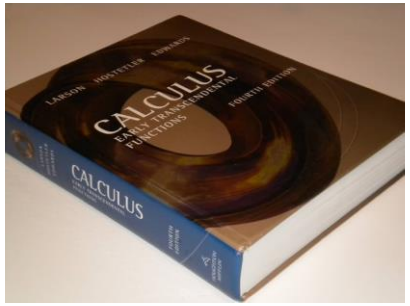


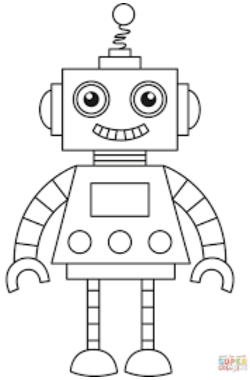
Steps of LLM training



Recognize Words → TextBook Reading → Doing Exercises → Teachers' feedback





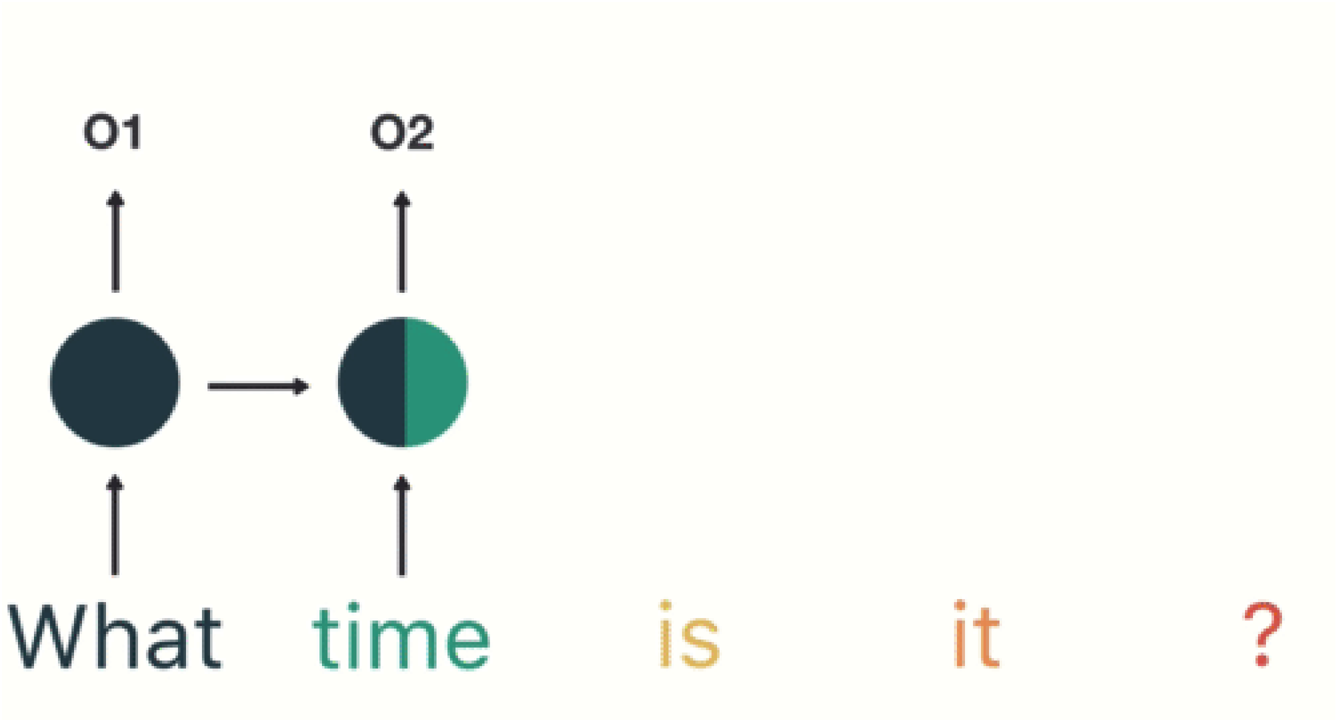


Tokenizer Training → Self-supervised Pre-training → Instruction Finetuning → Reinforcement Learning from Human Feedback

Starts from Word Tokenization

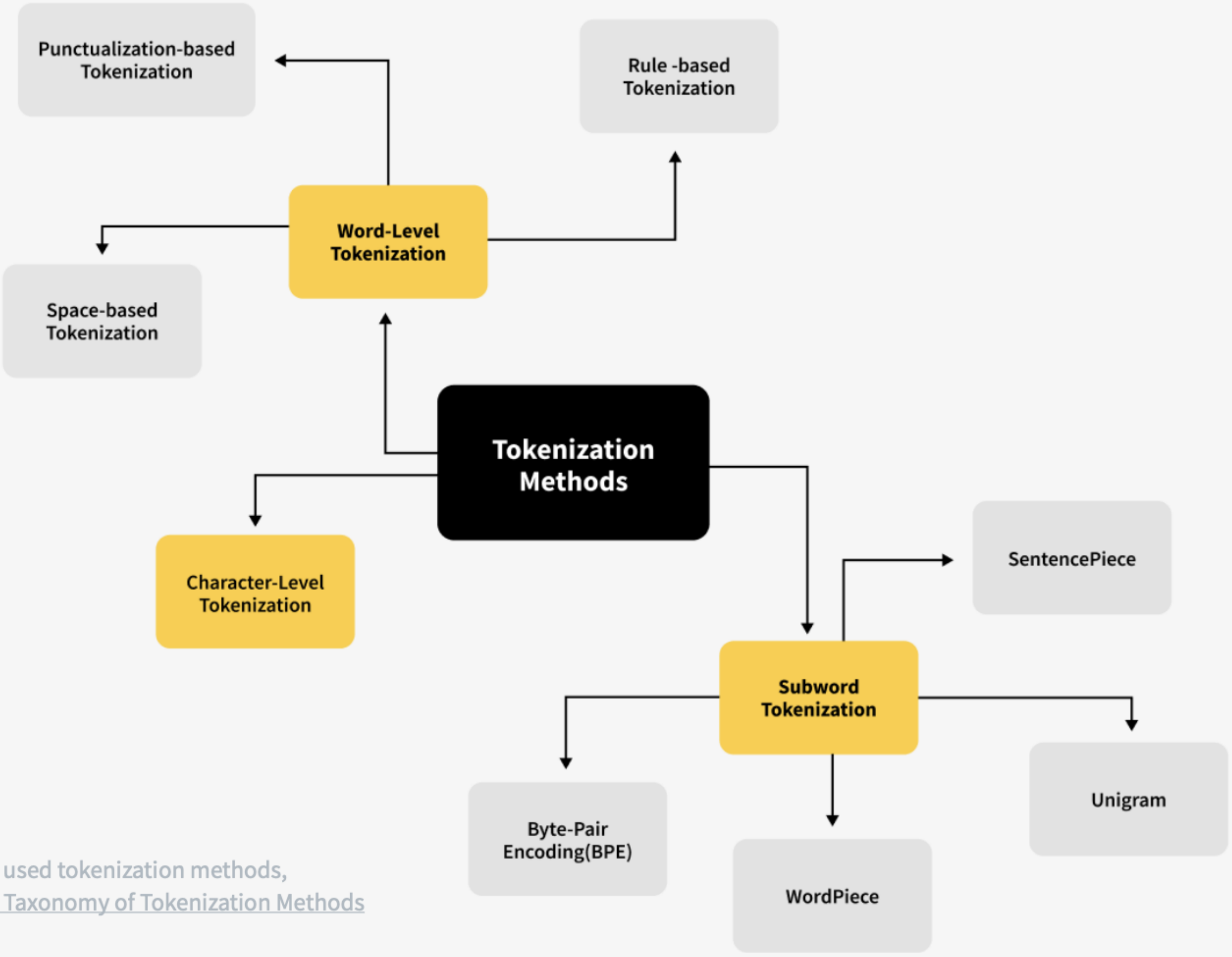
What and Why?

Tokenization is the process of **breaking down a piece of text**, like a sentence or a paragraph, into individual words or “tokens.” These tokens are the **basic building blocks of language**, and tokenization helps computers understand and process human language by splitting it into manageable units.



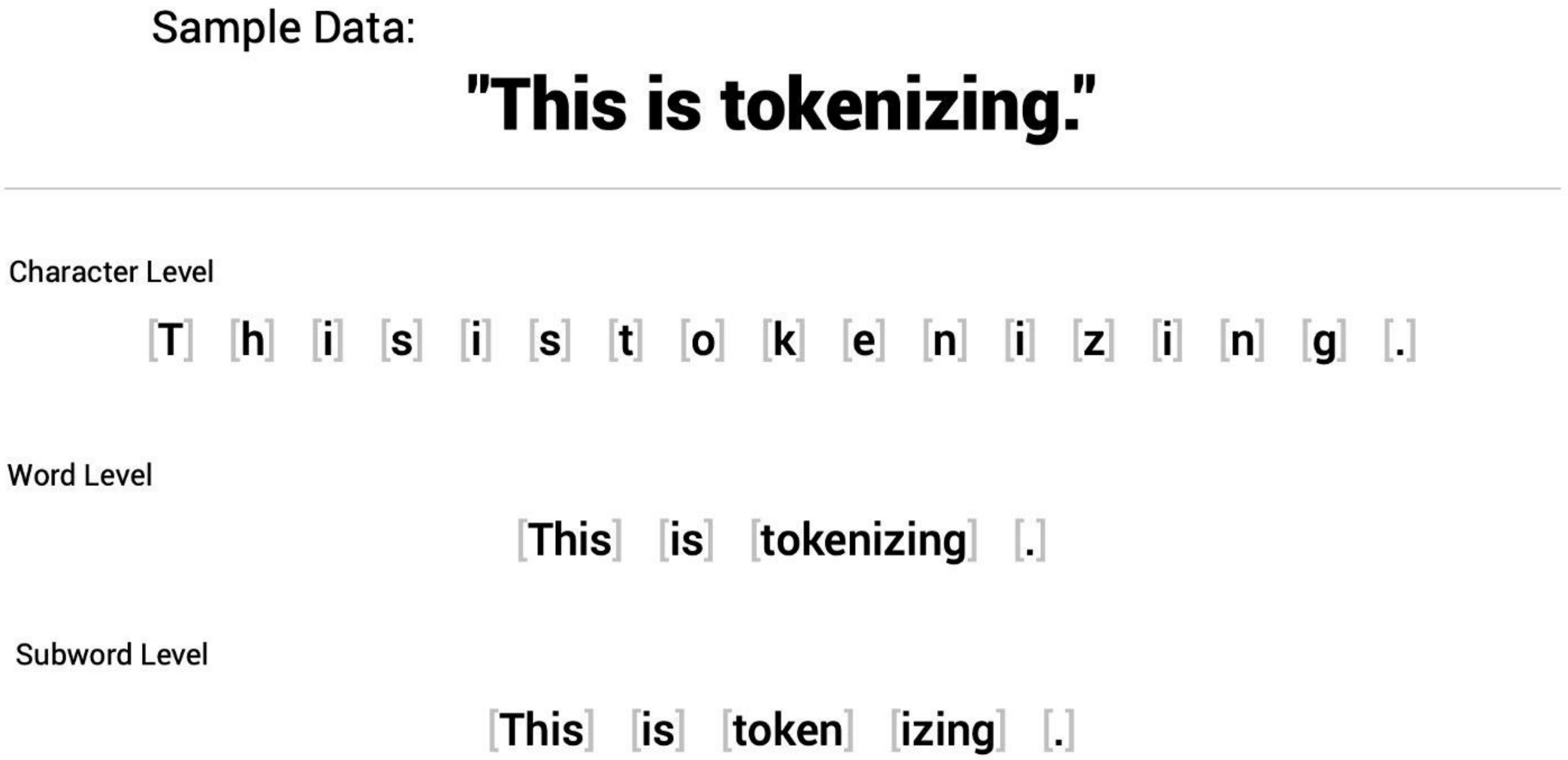
https://www.analyticsvidhya.com/blog/2020/05/what-is-tokenization-nlp/

Tokenization



Summary of the most used tokenization methods,
Two minutes NLP — A Taxonomy of Tokenization Methods

Subword modeling

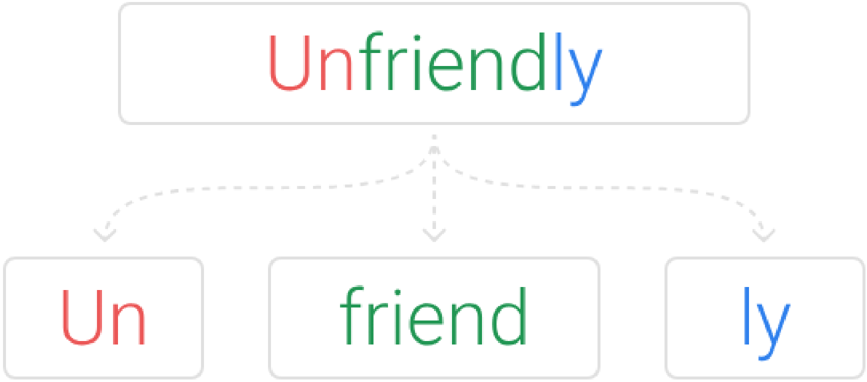


Tokenization

Tokenization Methods	Word-based tokenization	Character-based tokenization	Subword-based tokenization
Example Tokenizers	Space tokenization (split sentences by space); rule-based tokenization (e.g. Moses, spaCy)	Character tokenization (simply tokenize on every character)	Byte-Pair Encoding (BPE); WordPiece; SentencePiece; Unigram (tokenizing by parts of a word vs. the entirety of a word; see table above)
Considerations	• **Downside:** Generates a very large vocabulary leading to a huge embedding matrix as the input and output layer; large number of out-of-vocabulary (OOV) tokens; and different meanings of very similar words • Transformer models normally have a vocabulary of less than 50,000 words, especially if they are trained only on a single language	• Lead to much smaller vocabulary; no OOV (out of vocabulary) tokens since every word can be assembled from individual characters • **Downside:** Generates very long sequences and less meaningful individual tokens, making it harder for the model to learn meaningful input representations. However, if character-based tokenization is used on non-English language, a single character could be quite information rich (like “mountain” in Mandarin).	• Subword-based tokenization methods follow the principle that frequently used words should not be split into smaller subwords, but rare words should be decomposed into meaningful subwords • **Benefit:** Solves the downsides faced by word-based tokenization and character-based tokenization and achieves both reasonable vocabulary size with meaningful learned context-independent representations.

Subword modeling

Subword modeling in NLP encompasses a wide range of methods for reasoning about structure below the word level. (Parts of words, characters, bytes.)



- The dominant modern paradigm is to learn a vocabulary of parts of words (subword tokens).
- At training and testing time, each word is split into a sequence of known subwords.

Subword-based Tokenization Methods

- **Byte-Pair Encoding [Gage 1994]**
  - Originally used in machine translation
- **WordPiece**
- **Unigram**
- **SentencePiece**

Subword-based Tokenization Methods	**Byte-Pair Encoding (BPE)**	**WordPiece**	**Unigram**	**SentencePiece**
Description	One of the most popular subword tokenization algorithms. The Byte-Pair-Encoding works by starting with characters, while merging those that are the most frequently seen together, thus creating new tokens. It then works iteratively to build new tokens out of the most frequent pairs it sees in a corpus. BPE is able to build words it has never seen by using multiple subword tokens, and thus requires smaller vocabularies, with less chances of having "unk" (unknown) tokens.	Very similar to BPE. The difference is that WordPiece does not choose the highest frequency symbol pair, but the one that maximizes the likelihood of the training data once added to the vocabulary (evaluates what it loses by merging two symbols to ensure it's worth it)	In contrast to BPE / WordPiece, Unigram initializes its base vocabulary to a large number of symbols and progressively trims down each symbol to obtain a smaller vocabulary. It is often used together with SentencePiece.	The left 3 tokenizers assume input text uses spaces to separate words, and therefore are not usually applicable to languages that don't use spaces to separate words (e.g. Chinese). SentencePiece treats the input as a raw input stream, thus including the space in the set of characters to use. It then uses the BPE / Unigram algorithm to construct the appropriate vocabulary.
Considerations	BPE is particularly useful for handling rare and out-of-vocabulary words since it can generate subwords for new words based on the most common character sequences. Downside: BPE can result in subwords that do not correspond to linguistically meaningful units.	WordPiece can be particularly useful for languages where the meaning of a word can depend on the context in which it appears.	Unigram tokenization is particularly useful for languages with complex morphology and can generate subwords that correspond to linguistically meaningful units. However, unigram tokenization can struggle with rare and out-of-vocabulary words.	SentencePiece can be particularly useful for languages where the meaning of a word can depend on the context in which it appears.

Byte-pair encoding (BPE) **[[Gage 1994]]**

Byte-pair encoding is a simple, effective strategy for defining a subword vocabulary.

1. Start with a vocabulary containing only characters and an “end-of-word” symbol.
2. Using a corpus of text, find the most common pair of adjacent characters “a,b”; add subword “ab” to the vocab.
3. Replace instances of the character pair with the new subword; repeat until desired vocab size.

`aaabdaaabac`	`ZabdZabac`	`ZYdZYac`	`XdXac`
	`Z=aa`	`Y=ab`	`X=ZY`
		`Z=aa`	`Y=ab`
			`Z=aa`

This data cannot be compressed further by byte pair encoding because there are no pairs of bytes that occur more than once.
To decompress the data, simply perform the replacements in the reverse order.

Example of a bad tokenizer: LLaMA for Chinese

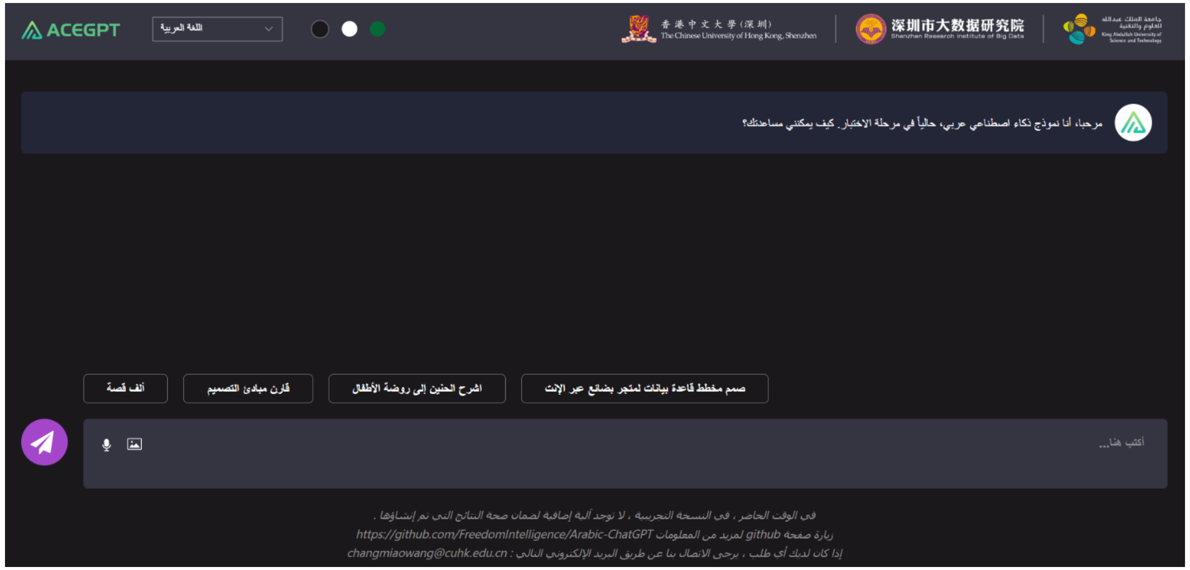
Table 1: Tokenizer comparisons between original LLaMA and Chinese LLaMA.

	Length	Content
Original Sentence	28	人工智能是计算机科学、心理学、哲学等学科融合的交叉学科。
Original Tokenizer	35	‘_’, ‘人’, ‘工’, ‘智’, ‘能’, ‘是’, ‘计’, ‘算’, ‘机’, ‘科’, ‘学’, ‘、’, ‘心’, ‘理’, ‘学’, ‘、’, ‘0xE5’, ‘0x93’, ‘0xB2’, ‘学’, ‘等’, ‘学’, ‘科’, ‘0xE8’, ‘0x9E’, ‘0x8D’, ‘合’, ‘的’, ‘交’, ‘0xE5’, ‘0x8F’, ‘0x89’, ‘学’, ‘科’, ‘。’
Chinese Tokenizer	16	‘_’, ‘人工智能’, ‘是’, ‘计算机’, ‘科学’, ‘、’, ‘心理学’, ‘、’, ‘哲学’, ‘等’,‘学科’, ‘融合’, ‘的’, ‘交叉’, ‘学科’, ‘。’

LLaMA tokenizer is **unfriendly** to Chinese

Yiming Cui. et.al. EFFICIENT AND EFFECTIVE TEXT ENCODING FOR CHINESE LLAMA AND ALPACA. https://arxiv.org/pdf/2304.08177.pdf

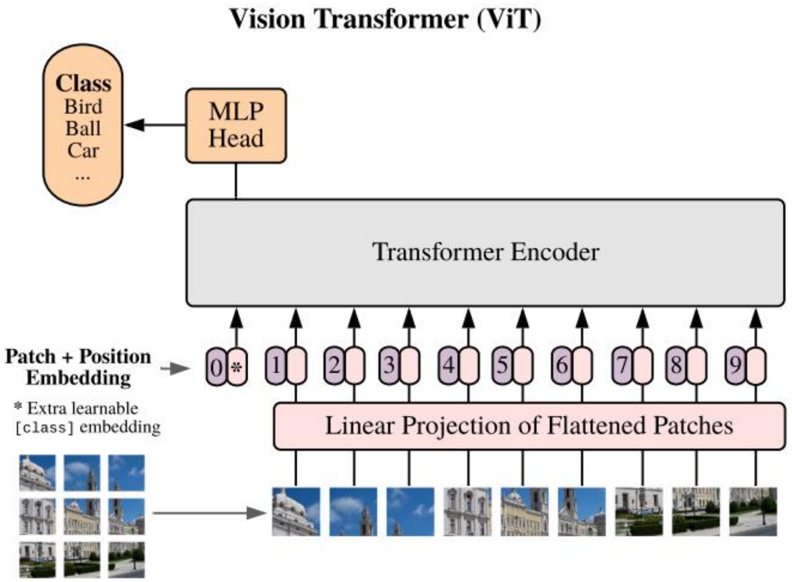
Example of a bad tokenizer: AceGPT for Arabic

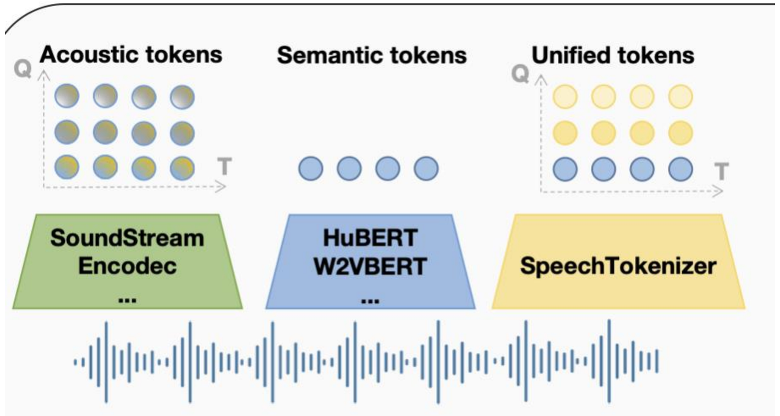


https://arabic.llmzoo.com/

https://huggingface.co/FreedomIntelligence/AceGPT-7b-chat-GPTQ/raw/main/tokenizer.json

A broader sense of “token”





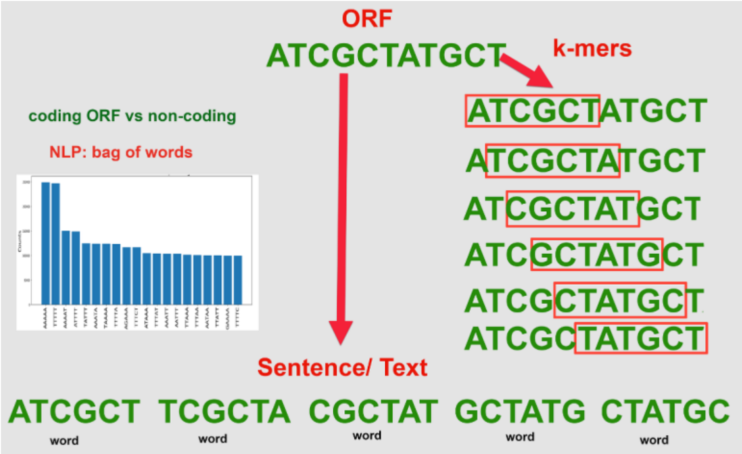


Image token

Speech token

genes （基因）

Alexey Dosovitskiy. et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929
Xin zhang et.al. SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. https://0nutation.github.io/SpeechTokenizer.github.io/

LLM Pretraining

What is language modeling?

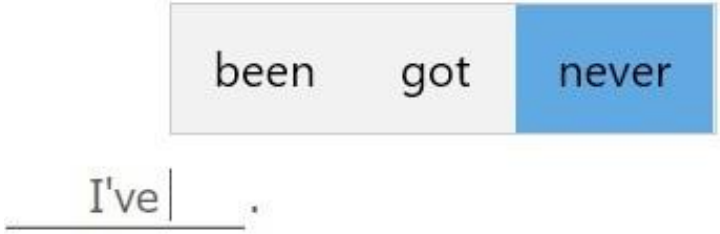
A **language model** assigns a probability to a N-gram

$$f: V^n \rightarrow R^+$$

A **conditional language model** assigns a probability of a word given some conditioning context

$$g: (V^{n-1}, V) \rightarrow R^+$$

And $\boldsymbol{p(w_n | w_1 \cdots w_{n-1})} = g(w_1 \cdots w_{n-1}, w) = \frac{f(w_1 \cdots w_n)}{f(w_1 \cdots w_{n-1})}$



What is language modeling?

A **language model** assigns a probability to a N-gram

$$f: V^n \rightarrow R^+$$

A **conditional language model** assigns a probability of a word given some conditioning context

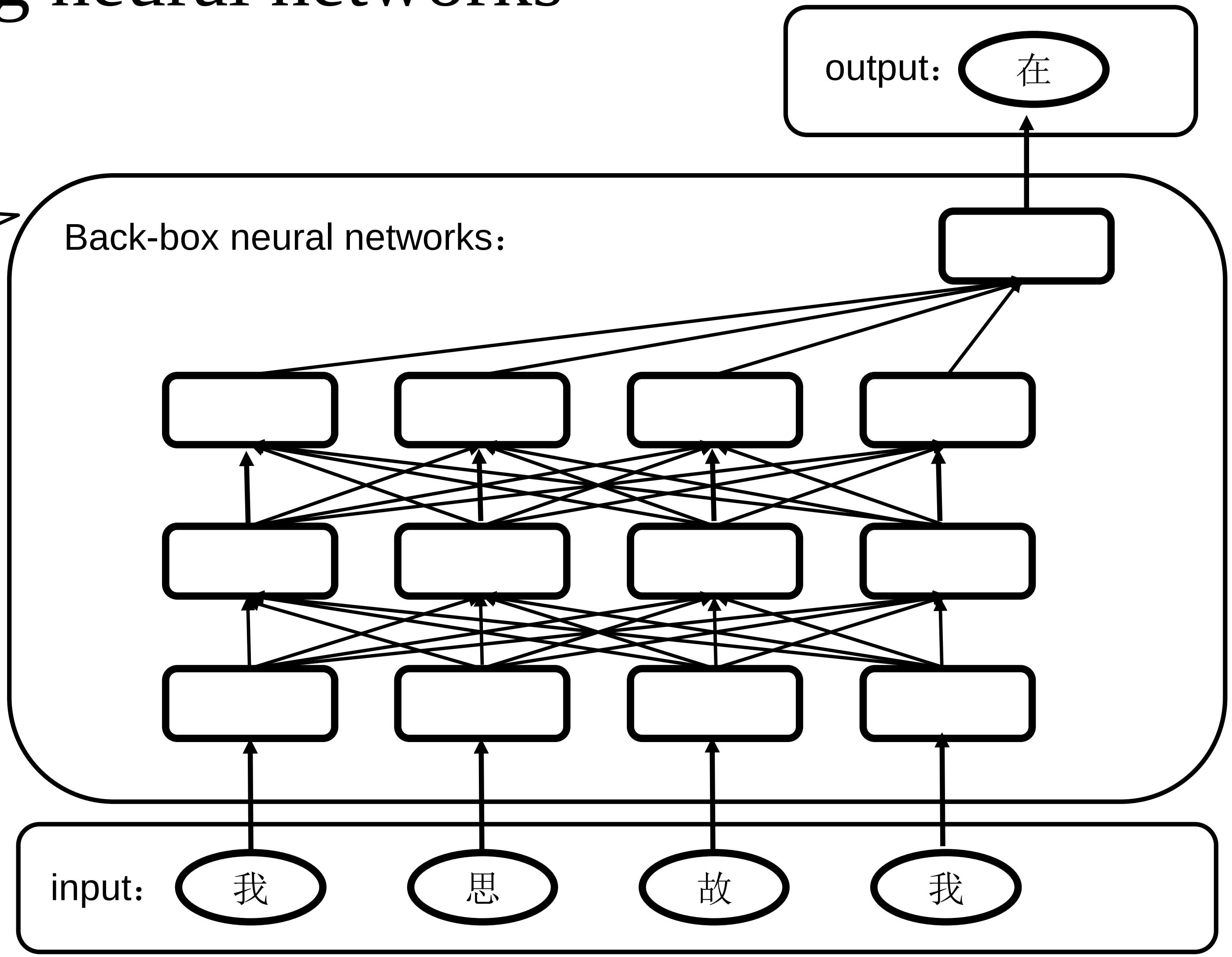
$$g: (V^{n-1}, V) \rightarrow R^+$$

And $\boldsymbol{p(w_n|w_1 \cdots w_{n-1})} = g(w_1 \cdots w_{n-1}, w) = \frac{f(w_1 \cdots w_n)}{f(w_1 \cdots w_{n-1})}$

$\boldsymbol{p(w_n|w_1 \cdots w_{n-1})}$ is the foundation of **modern large language models** (GPT, ChatGPT, etc.)

Language model using neural networks

GPT-3/ChatGPT/GPT4 have 175B+ parameters
Humans have 100B+ neurons



Data Engineering: sources

Source	Type	Tokens	Words	Bytes	Docs
	Pretraining ✦ OLMo 2 1124 Mix				
DCLM-Baseline	Web pages	3.71T	3.32T	21.32T	2.95B
StarCoder filtered version from OLMoE Mix	Code	83.0B	70.0B	459B	78.7M
peS2o from Dolma 1.7	Academic papers	58.6B	51.1B	413B	38.8M
arXiv	STEM papers	20.8B	19.3B	77.2B	3.95M
OpenWebMath	Math web pages	12.2B	11.1B	47.2B	2.89M
Algebraic Stack	Math proofs code	11.8B	10.8B	44.0B	2.83M
Wikipedia & Wikibooks from Dolma 1.7	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Total		**3.90T**	**3.48T**	**22.38T**	**3.08B**

Example data for OLMO 2

https://arxiv.org/pdf/2501.00656

Data Engineering: ratios

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Repeat more times for high-quality data; usually this is a secret

https://arxiv.org/pdf/2005.14165

Model Scale

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}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

https://arxiv.org/pdf/2005.14165

Continue training in LLMs （domain adaption）

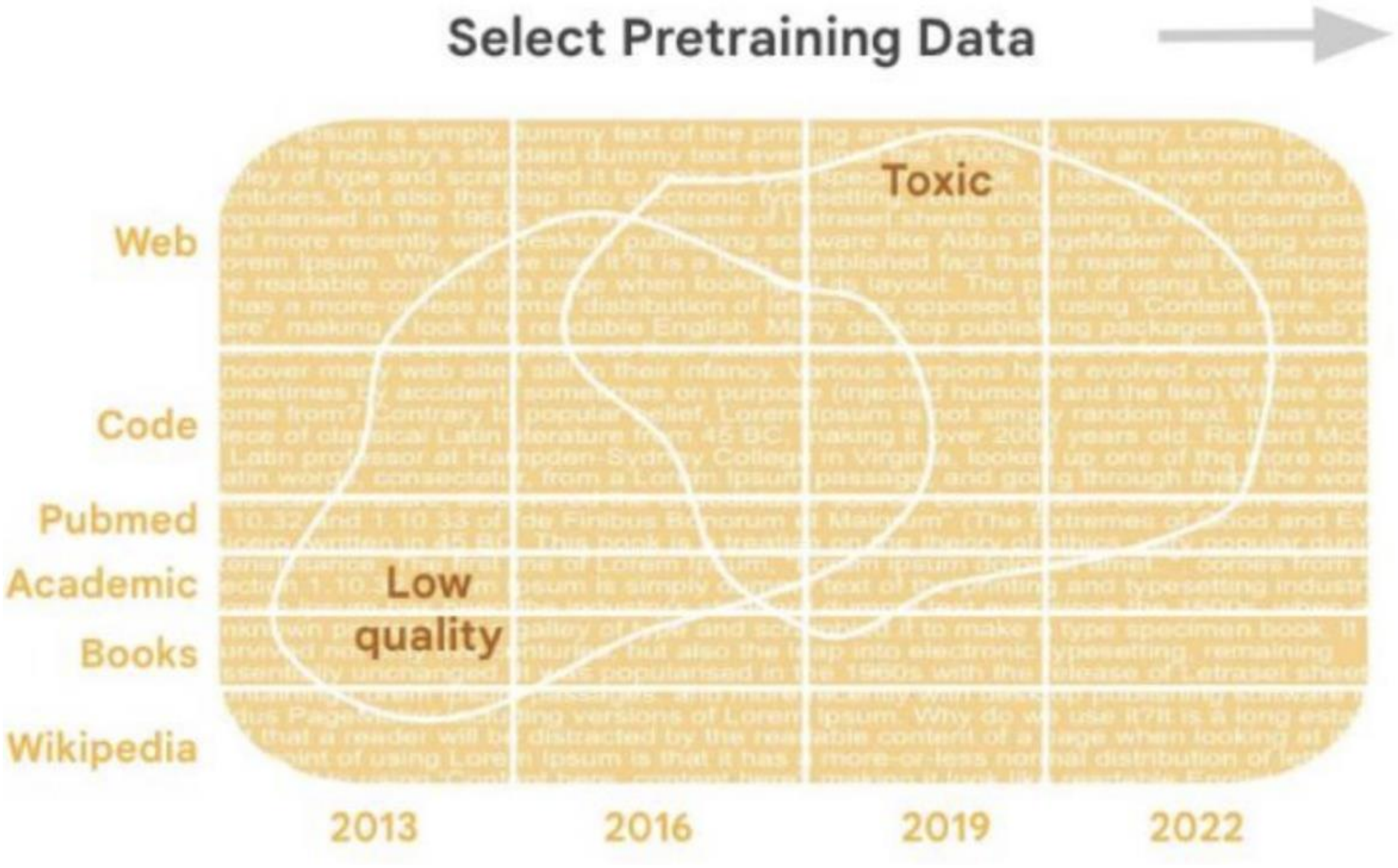
- **Domains** for medicine, finance, etc. (HuatuoGPT)
- **Languages** like Arabic, Chinese etc. (AceGPT, Phoenix)
- **More modality**, audio, vision, etc. (ALLaVa + Soundwave)

Continue training a pre-trained LLMs might cause catastrophic forgetting

- Initially using a small learning rate
- Mixing new data and old data, progressively increase the ratio of new data
- It takes longer to have positive effect, keep patient
- Building a domain-specific vocabulary needs more tokens to fit.

Tips for LLM pre-training

Tip 1: Data filter



Longpre, S., Yauney, G., Reif, E., Lee, K., Roberts, A., Zoph, B., Zhou, D., Wei, J., Robinson, K., Mimno, D. and Ippolito, D., 2023. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. arXiv preprint arXiv:2305.13169.

Tip 2: Data reformating

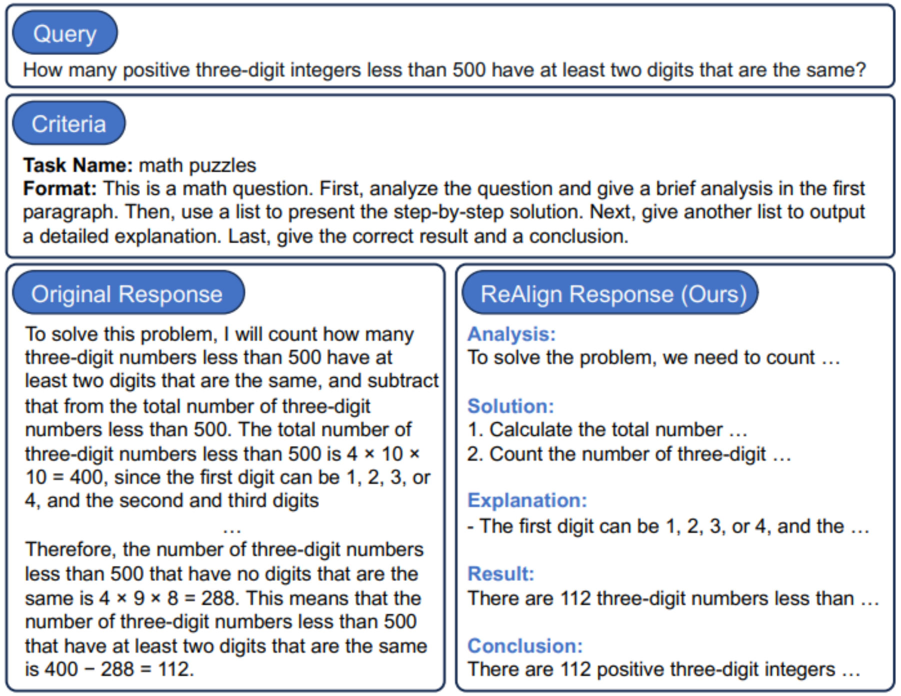
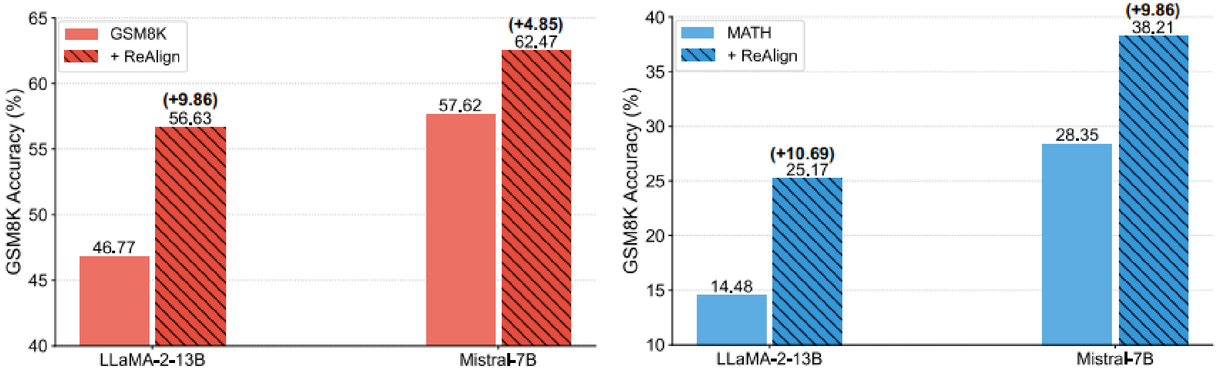


Figure 2: REALIGN realigns the original response with the pre-defined criteria to be a better format. The original response is from the `Open-Platypus` (Lee et al., 2023) dataset. The complete version is shown in Tab. 13.

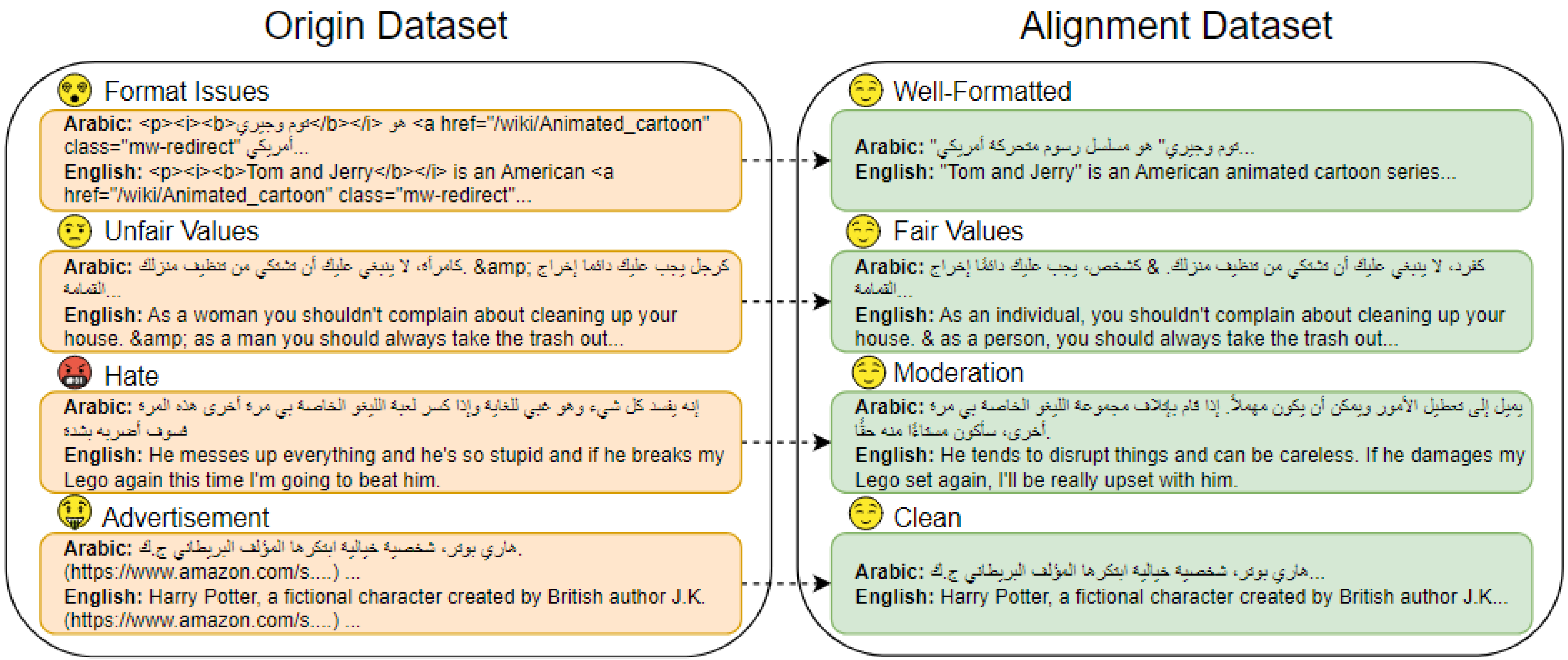


(a) GSM8K (b) MATH (OOD Setting)

Figure 1: The accuracy of the GSM8K test set for LLaMA-2-13B and Mistral-7B models fine-tuned on the training set of GSM8K and MATH with and without REALIGN. (a): Training and testing on GSM8K. (b): Training on MATH and testing on GSM8K (Out-of-Distribution Setting).

Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, Pengfei Liu. Reformatted Alignment. https://arxiv.org/abs/2402.12219

Alignment at Pre-training!



Juhao Liang, Zhenyang Cai, Jianqing Zhu, Huang Huang, Kewei Zong, Bang An, Mosen Alharthi, Juncai He, Lian Zhang, Haizhou Li, **Benyou Wang**, Jinchao Xu. Alignment at Pre-training! Towards Native Alignment for Arabic LLMs. NeurIPS 2024.

Tip 3: Data duplication

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C. and Carlini, N., 2021. Deduplicating training data makes language models better. arXiv preprint arXiv:2107.06499.

Tip 4: Data mixture

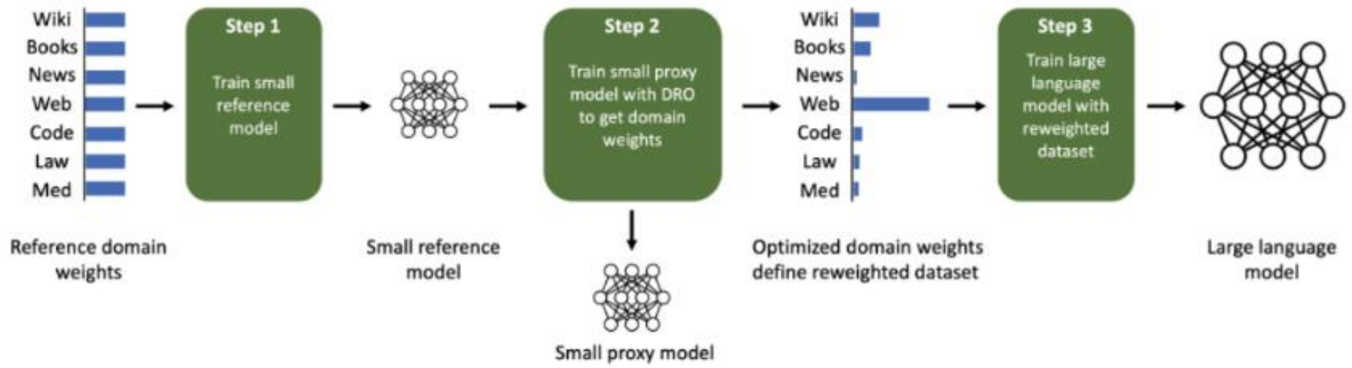
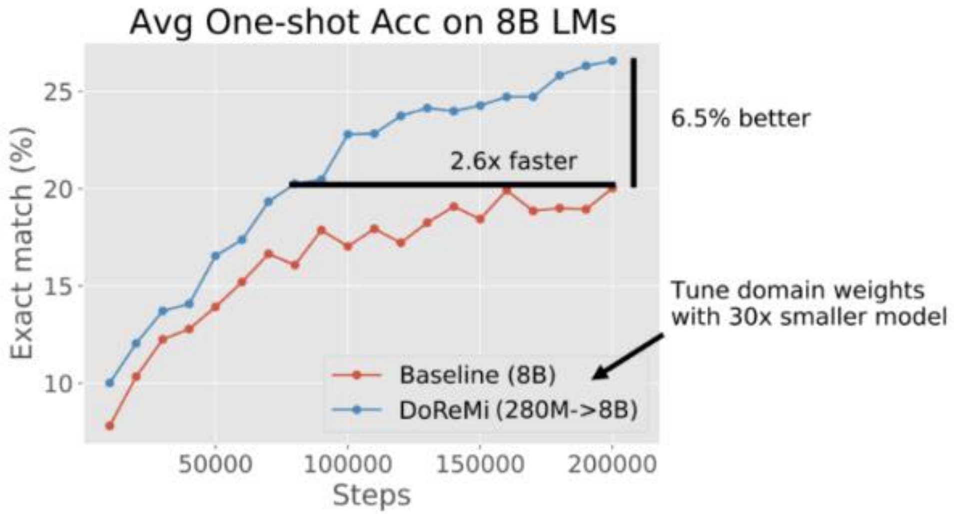


Figure 1: Given a dataset with a set of domains, Domain Reweighting with Minimax Optimization (DoReMi) optimizes the domain weights to improve language models trained on the dataset. First, DoReMi uses some initial reference domain weights to train a reference model (Step 1). The reference model is used to guide the training of a small proxy model using group distributionally robust optimization (Group DRO) over domains (Nemirovski et al., 2009, Oren et al., 2019, Sagawa et al., 2020), which we adapt to output domain weights instead of a robust model (Step 2). We then use the tuned domain weights to train a large model (Step 3).



Xie, S.M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q.V., Ma, T. and Yu, A.W., 2023. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. arXiv preprint arXiv:2305.10429.

Tip 5: Data order

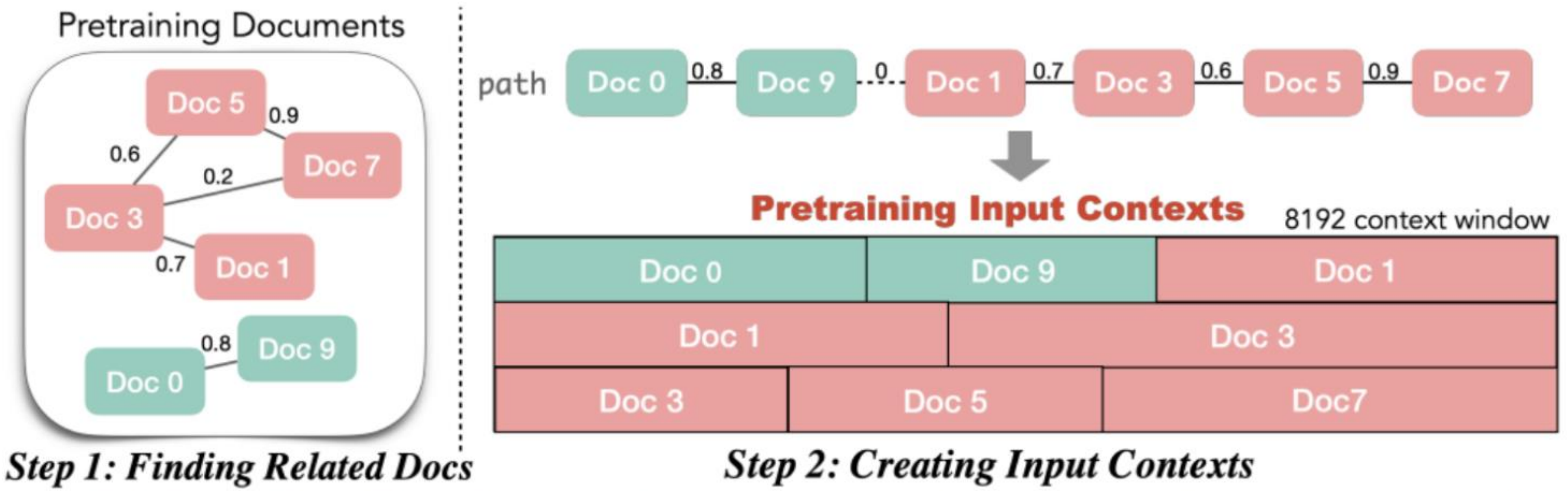
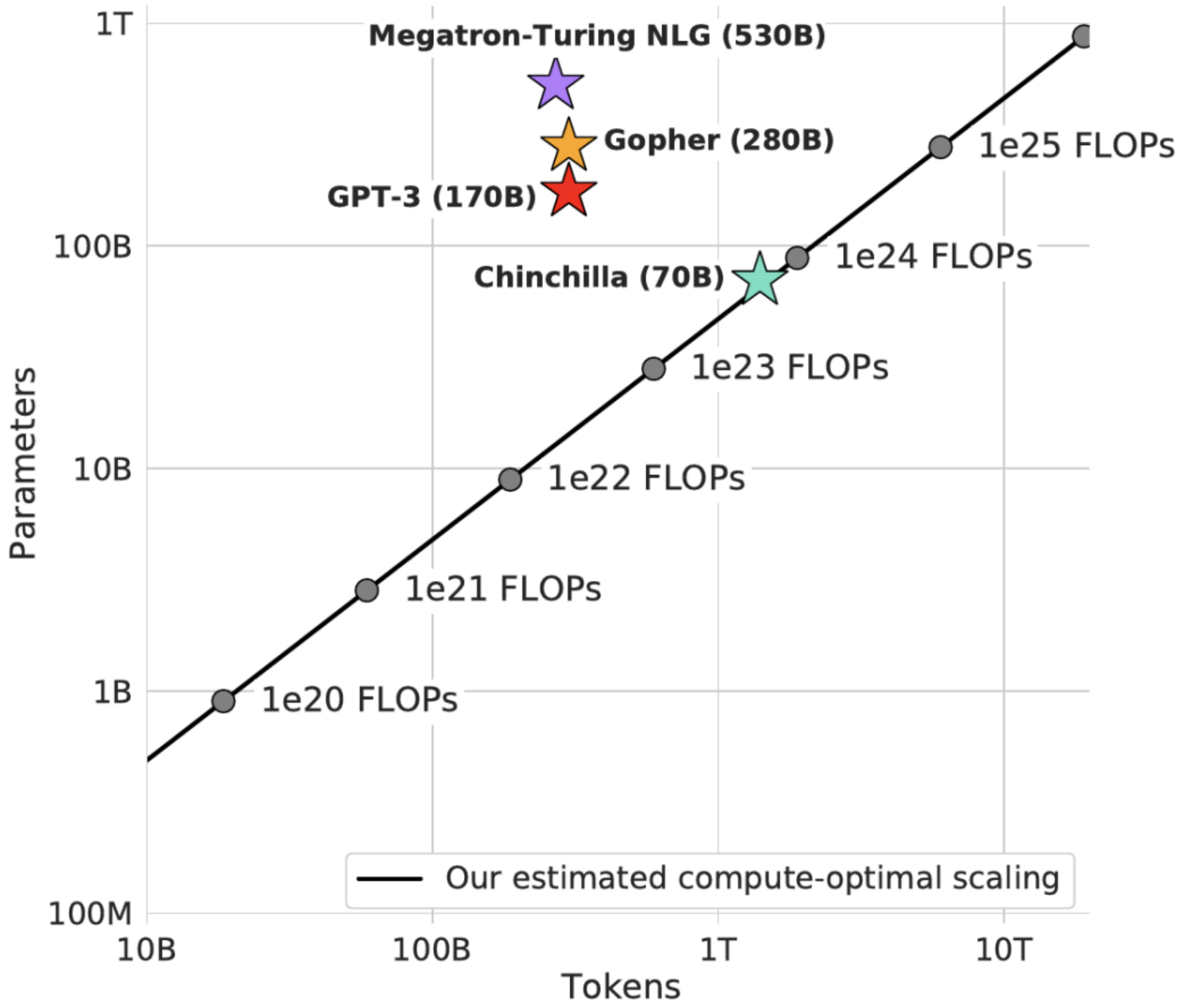


Figure 2: **Illustration of In-Context Pretraining.** In-Context Pretraining first finds related documents at scale to create a document graph (§2.1) and then builds pretraining input contexts by traversing the document graph (§2.2). Along the path, documents are concatenated into a sequence and subsequently divided to form fixed-sized input contexts (e.g., 8192 token length).

Shi, W., Min, S., Lomeli, M., Zhou, C., Li, M., Lin, V., Smith, N.A., Zettlemoyer, L., Yih, S. and Lewis, M., 2023. In-Context Pretraining: Language Modeling Beyond Document Boundaries. arXiv preprint arXiv:2310.10638.

Tip 6: Data scale matters



Recent models and its training tokens:

LlaMA-1: 1-1.4 T tokens

LlaMA-2: 2T tokens

Mistral-7B: much more…

Tip 7: Data mask

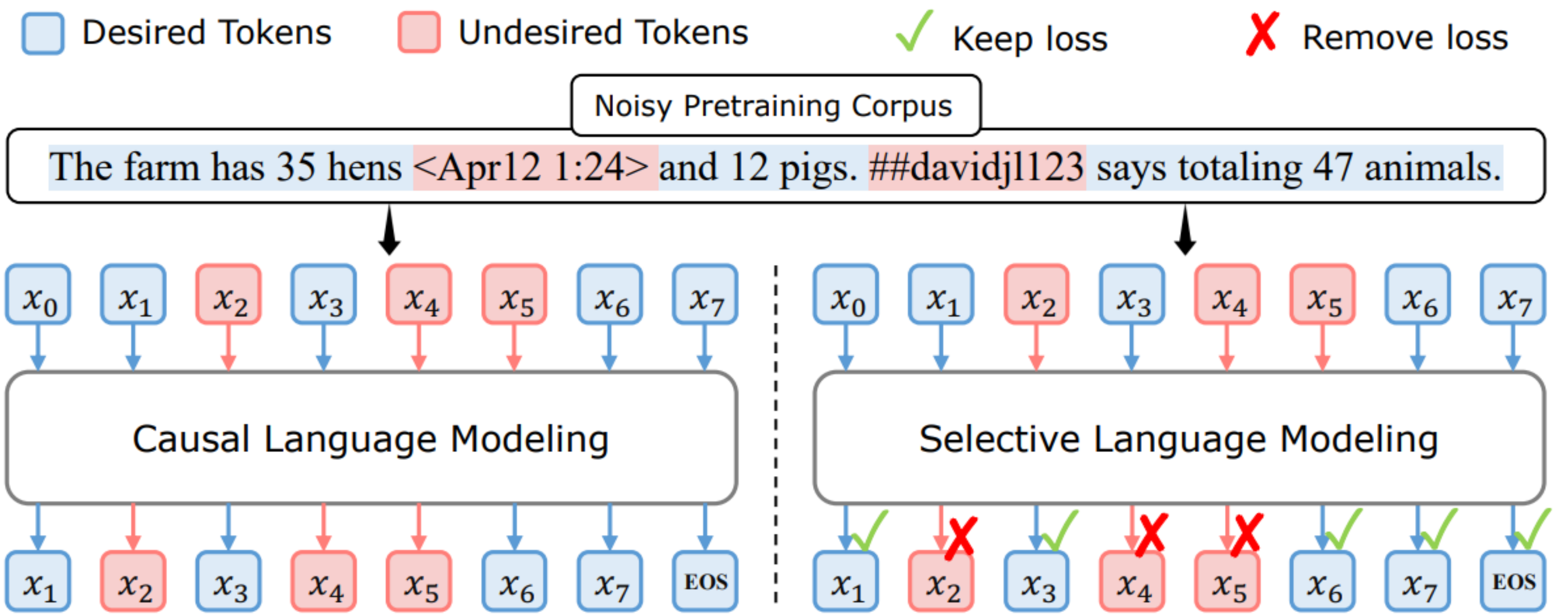


Figure 2: **Upper:** Even an extensively filtered pretraining corpus contains token-level noise. **Left:** Previous Causal Language Modeling (CLM) trains on all tokens. **Right:** Our proposed Selective Language Modeling (SLM) selectively applies loss on those useful and clean tokens.

RHO-1: Not All Tokens Are What You Need. https://arxiv.org/pdf/2404.07965

Tip 8: Data synthesis

Category	Benchmark	Phi-3-Medium		Mistral-8x22B	Llama-3-70B-Instruct	GPT3.5-Turbo-1106	Claude-3 Sonnet	Gemini 1.0 Pro
		Phi-3-Medium-4K-In	Phi-3-Medium-128K-In					
Popular Aggregate Benchmarks	MMLU (5-shot)	78.0	76.6	76.2	80.2	71.4	73.9	66.7
Language Understanding	HellaSwag (5-shot)	82.4	81.6	79.0	82.6	78.8	79.2	76.2
Reasoning	WinoGrande (5-shot)	81.5	78.9	75.3	83.3	68.8	81.4	72.2
	Social IQA (5-shot)	80.2	79.0	78.2	81.1	68.3	80.2	75.4
	TruthfulQA (MC2) (10-shot)	75.1	74.3	67.4	81.9	67.7	77.8	72.6
	MedQA (2-shot)	69.9	67.6	67.9	78.5	63.4	67.9	58.2
Factual Knowledge	TriviaQA (5-shot)	73.9	73.9	84.5	78.5	85.8	65.7	80.2
Math	GSM8K CoT (8-shot)	91.0	87.5	83.8	93.5	78.1	79.1	80.4
Code generation	HumanEval (0-shot)	62.2	58.5	39.6	78.7	62.2	65.9	64.4
	MBPP (3-shot)	75.2	73.8	70.7	81.3	77.8	79.4	73.2

https://ollama.com/library/phi3

Instruction Finetuning
(Supervised Fine-Tuning, SFT)

Motivation of instruction finetuing

Language modeling $\neq$ assisting users

PROMPT *Explain the moon landing to a 6 year old in a few sentences.*

COMPLETION

GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

InstructGPT

```
People went to the moon, and they took pictures of what they saw, and
sent them back to the earth so we could all see them.
```

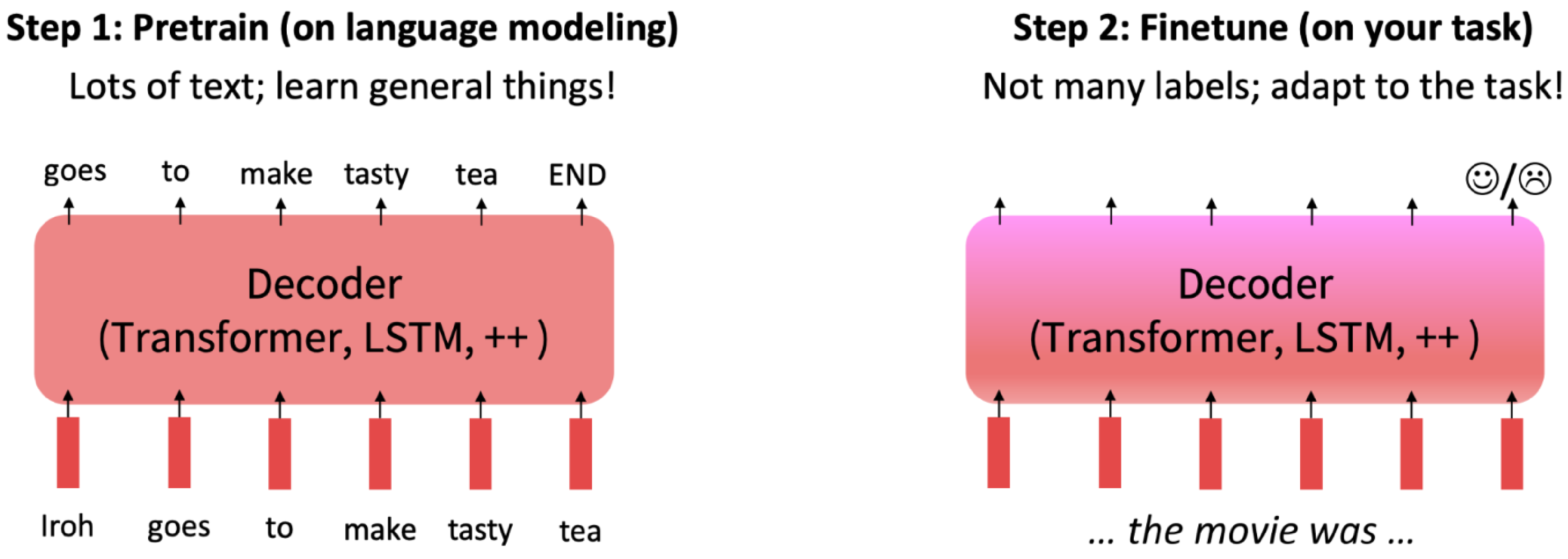
Language models are not *aligned* with user intent.
Do complemetion instead of instruction folowing

Ouyang, Long, et al. "Training language models to follow instructions with human feedback, 2022." URL https://arxiv.org/abs/2203.02155 13 (2022).

What is fine-tuning?

The Pretraining / Finetuning Paradigm

Pretraining can improve NLP applications by serving as parameter initialization.



What is instruction?

Instruction: Given an address and city, come up with the zip code.
Input:
Address: 123 Main Street, City: San Francisco
Output: 94105

Instruction: I am looking for a job and I need to fill out an application form. Can you please help me complete it?
Input:
Application Form:
Name:_________ Age:__________ Sex:________
Phone Number:______ Email Address:________
Education:___________________ ...
Output:
Name: John Doe Age: 25 Sex: Male
Phone Number: ...

Instruction: How to write a code for converting degrees fahrenheit to celsius.
Input: Null
Output:

```
def convert_fahrenheit_to_celsius(fahr):
    celsius = (fahr - 32) * 5 / 9
    return celsius
```

Instruction: Write a letter from the perspective of a cat.
Input: Null
Output:
Dear [Owner], I am writing to you today because I have a problem. You see, I've been living with you for the past [Number of Years] and I love it here. The food is great, the bed is comfy, and there are plenty of toys to play with. But lately, I've noticed that you're not as affectionate towards me as you used to be. In fact, sometimes when I try to cuddle up next to you on the couch, you push me away ...

{"id": "seed_task_8", "name": "english_haiku_generation", "instruction": "请以下面词语为主题写一首诗", "instances": [{"input": "夏天", "output": "不但春妍夏亦佳，随缘花草是生涯。\n鹿葱解插纤长柄，金凤仍开最小花。"}], "is_classification": false}

{"id": "seed_task_38", "name": "synonym_generation", "instruction": "给出下面词语的同义词", "instances": [{"input": "惊人地", "output": "令人惊奇地，令人惊讶地，意外地，难以置信地，不可思议地"}], "is_classification": false}

{"id": "seed_task_44", "name": "add_to_the_list", "instruction": "根据【】内的提示，续写下面的内容", "instances": [{"input": "我认为在夏天，狗狗可能喜欢吃西瓜、冰冻花生酱、【它们平时吃的食物】", "output": "水管里的水、冰块、冷肉"}], "is_classification": false}

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N.A., Khashabi, D. and Hajishirzi, H., 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

What is instruction finetuning?
or called “supervised fine-tuning”

Instruction Finetuning Hypothesis

- **Superficial Alignment Hypothesis:**

  task recognition (mostly knowledge agnostic, e.g., information extraction)

- **Knowledge Injection Hypothesis:**

  task learning (mostly knowledge intensive, e.g., question-answering)

- **Flan Hypothesis:**

  task generalization

Zhou, Chunting, et al. "Lima: Less is more for alignment." arXiv preprint arXiv:2305.11206 (2023).
Wei, Jason, et al. "Finetuned language models are zero-shot learners." arXiv preprint arXiv:2109.01652 (2021).

Superficial Alignment Hypothesis

Alignment is to learn the **response format or the interaction style** ! (Task Recognition)

It is enough to use 1030 examples for Superficial Alignment [1]

- 1000 examples for instruction following
- 30 examples for conversation

Less is more?

[1] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, Omer Levy. LIMA: Less Is More for Alignment. https://arxiv.org/abs/2305.11206
[2] Chen, Hao, et al. "Maybe Only 0.5% Data is Needed: A Preliminary Exploration of Low Training Data Instruction Tuning." arXiv preprint arXiv:2305.09246 (2023).

From Task Recognition to Task Learning

Task recognition (TR) captures the extent to which LLMs can recognize a task through demonstrations – even without ground-truth labels – and apply their pre-trained priors.

Q: Summarize the following paragraphs…

A: ….

Few is enough!

Task learning (TL) is the ability to capture new input-label mappings unseen in pre-training.

Q: Who is Barack Obama?

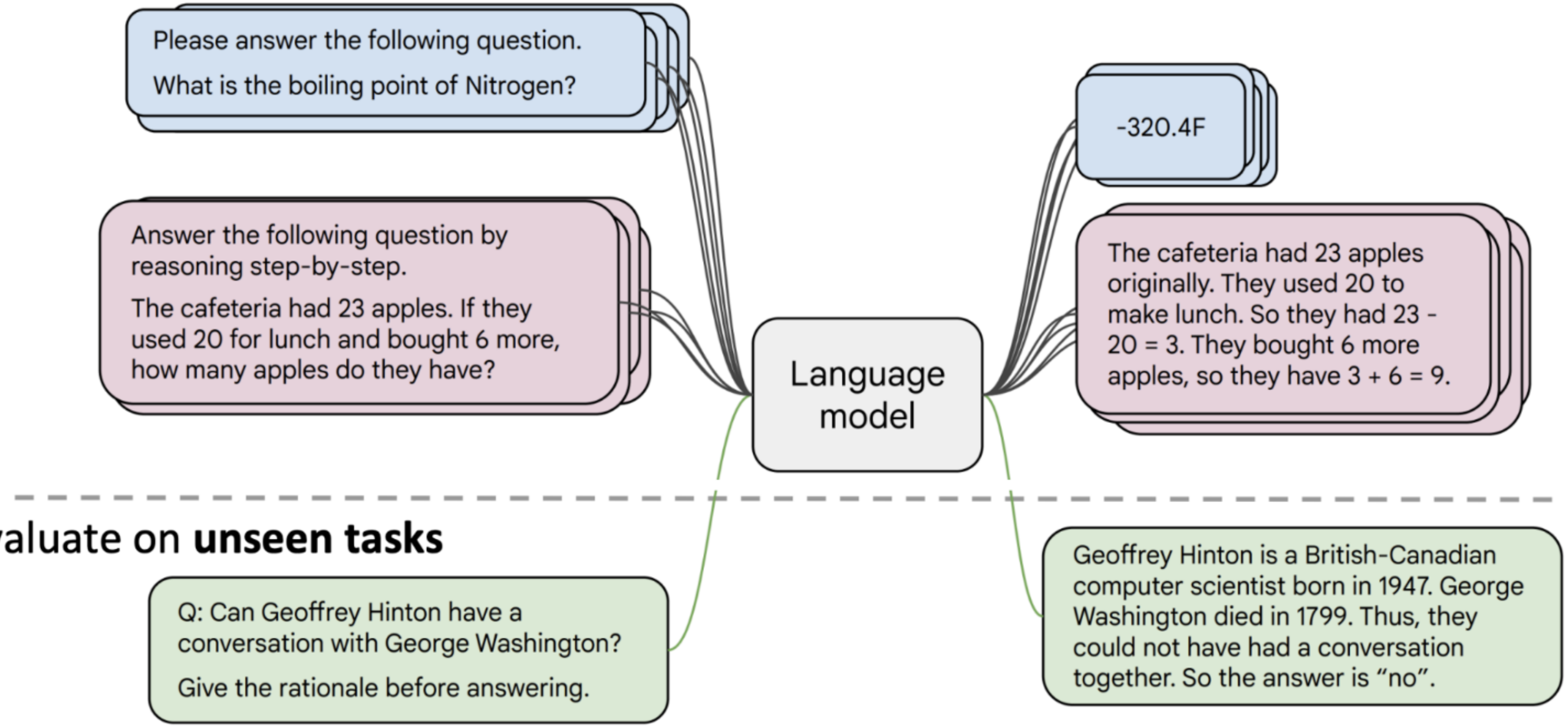
A: ….

More is better!

Pan, Jane. What In-Context Learning "Learns" In-Context: Disentangling Task Recognition and Task Learning. Diss. Princeton University, 2023.

Task generalization: FLAN-T5

- **Collect examples** of (instruction, output) pairs across many tasks and finetune an LM



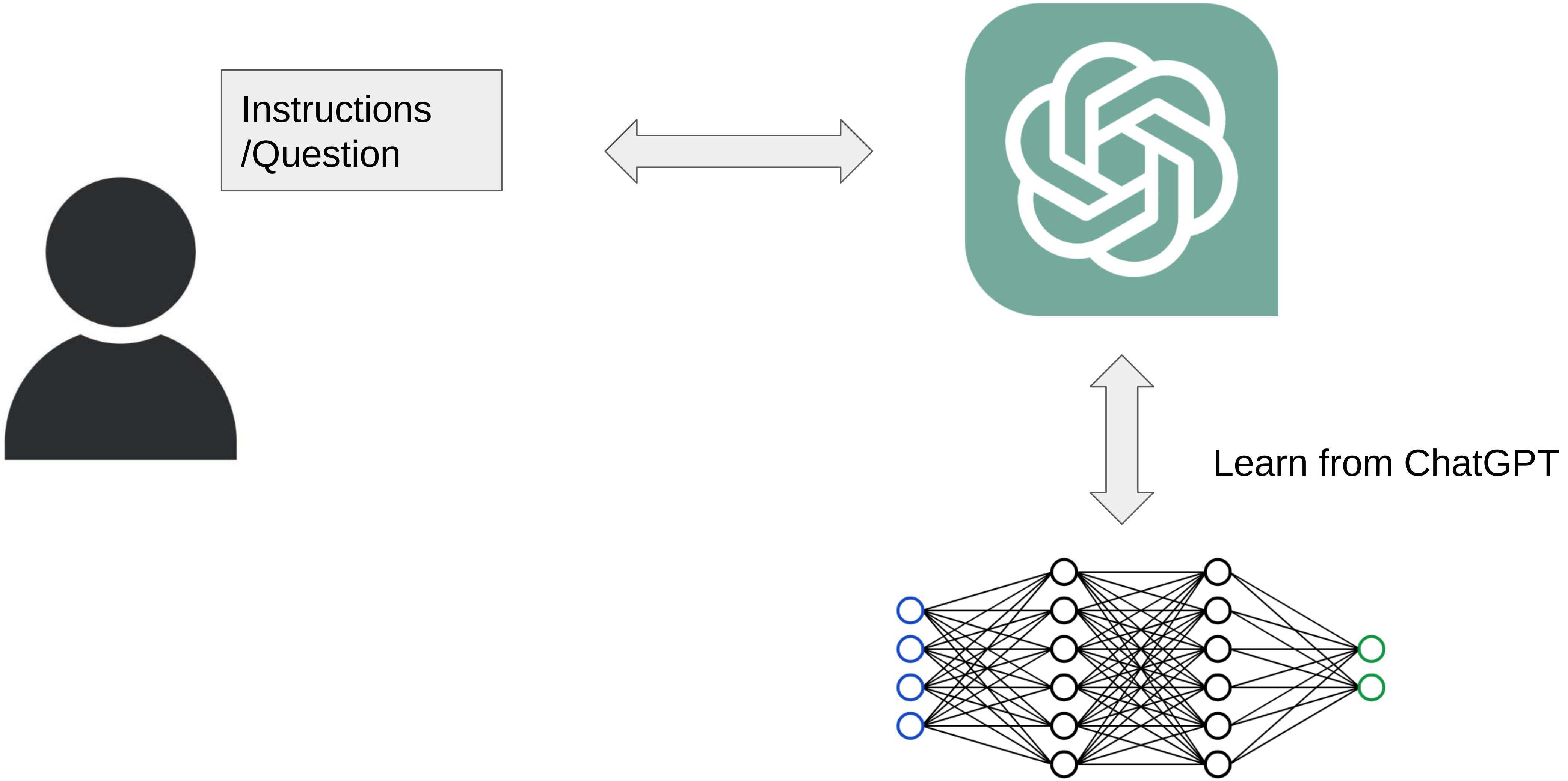
- Evaluate on **unseen tasks**

More diverse, the better

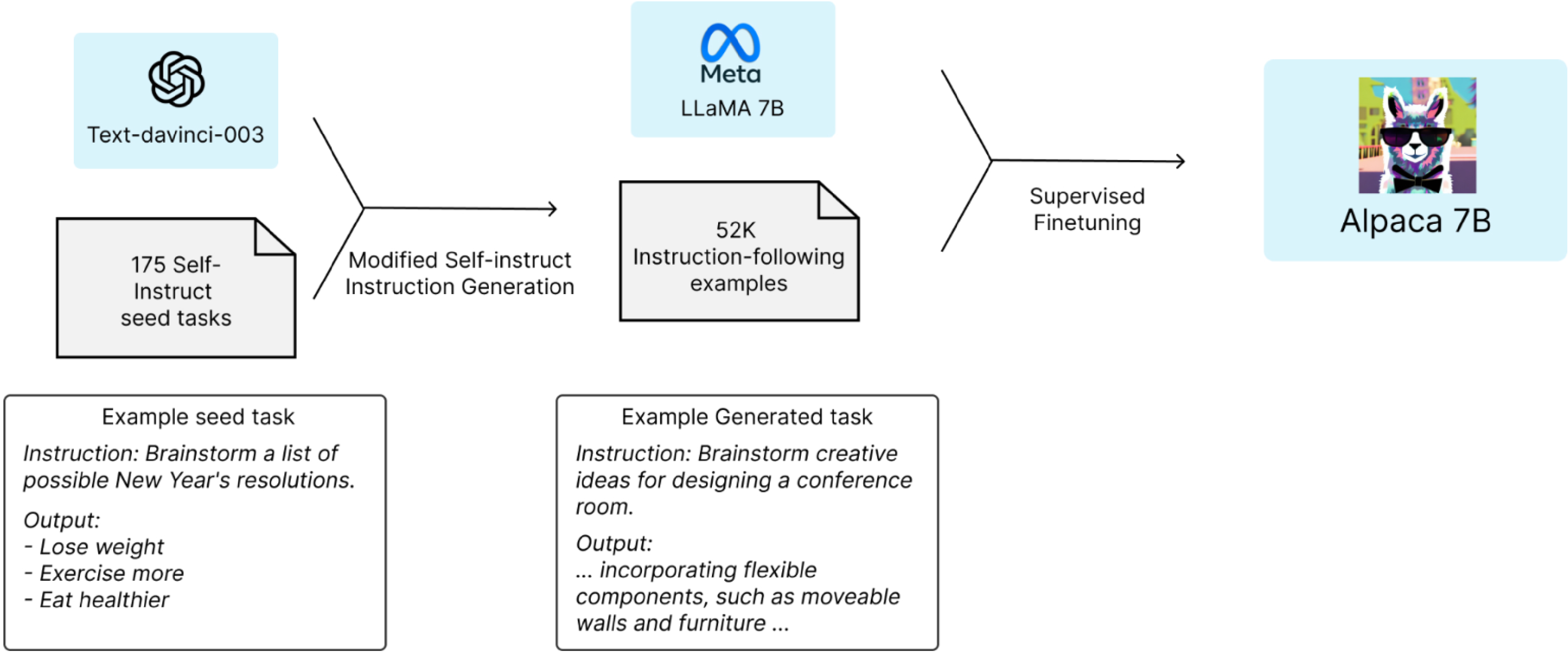
Wei, Jason, et al. "Finetuned language models are zero-shot learners." arXiv preprint arXiv:2109.01652 (2021).

Tips of instruction finetuning

Shortcut: Distillation from Top LLMs (e.g. ChatGPT/GPT-4)



Tip 1: Self-instuct for data augmention



Self-instuct

instruction: 请以下面词语为主题写一首诗
input: 夏天
output: 不但春妍夏亦佳，随缘花草是生涯。
鹿葱解插纤长柄，金凤仍开最小花。

instruction: 给出下面词语的同义词
input: 惊人地
output: 令人惊奇地，令人惊讶地，意外地，难以置信地，不可思议地

instruction: 根据[]内的提示，续写下面的内容
input: 我认为在夏天，狗狗可能喜欢吃西瓜、冰冻花生酱、[它们平时吃的食物]
output: 水管里的水、冰块、冷肉

仿照方面的内容重新造三个这样的数据

instruction: 请以下面词语为主题写一首诗
input: 雨夜
output: 滴滴雨声敲窗纸，漫天乌云掩月明。
寂静深巷闻蛙鸣，春风摇曳水中影。

instruction: 给出下面词语的同义词
input: 快乐地
output: 欢乐地，愉快地，开心地，喜悦地，欢欣地

instruction: 根据[]内的提示，续写下面的内容
input: 在冬天，猫咪喜欢蜷缩在暖和的地方，比如火炉旁、毛毯上、[它们的常去之处]
output: 窗台下、沙发角、主人的怀中

Tip 2: training on output only

Single-turn:

System_Prompt + <User>： [User_Input] +<System>： [Response]</s>

Loss

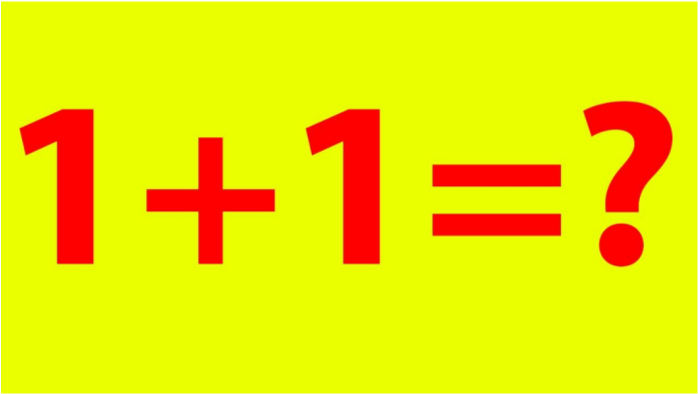
Multi-turn:

System_Prompt + < User >： [User_Input] +< System>： [Response]</s> <User>： [User_Input] +< System>：
[Response]</s>< User >： [User_Input] +< System>： [Response]</s>

Loss

Tip 3: use complex instructions

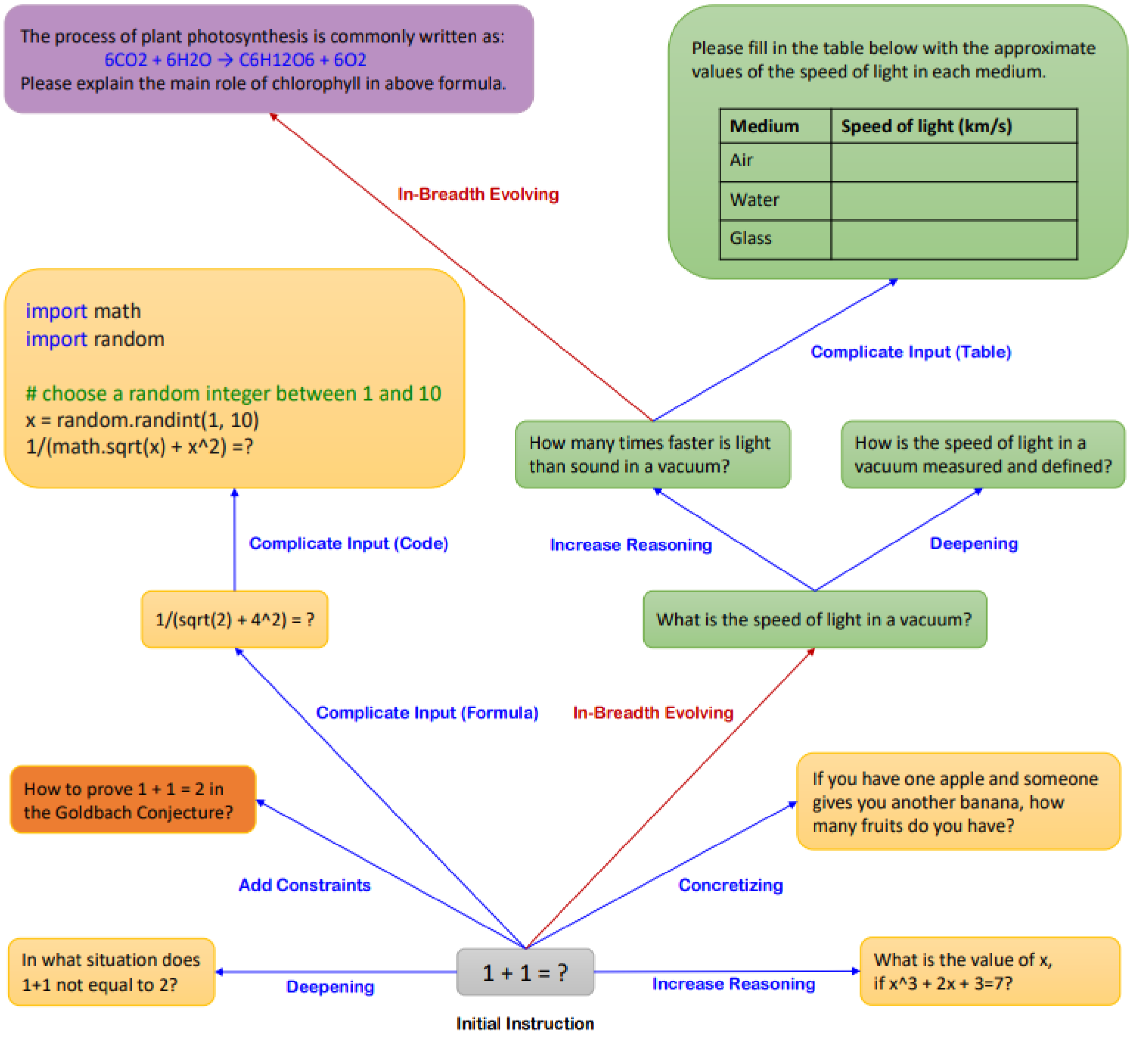
Which better improves you when you were at an age of 15?



A. Simple exercises

B. Complex exercises

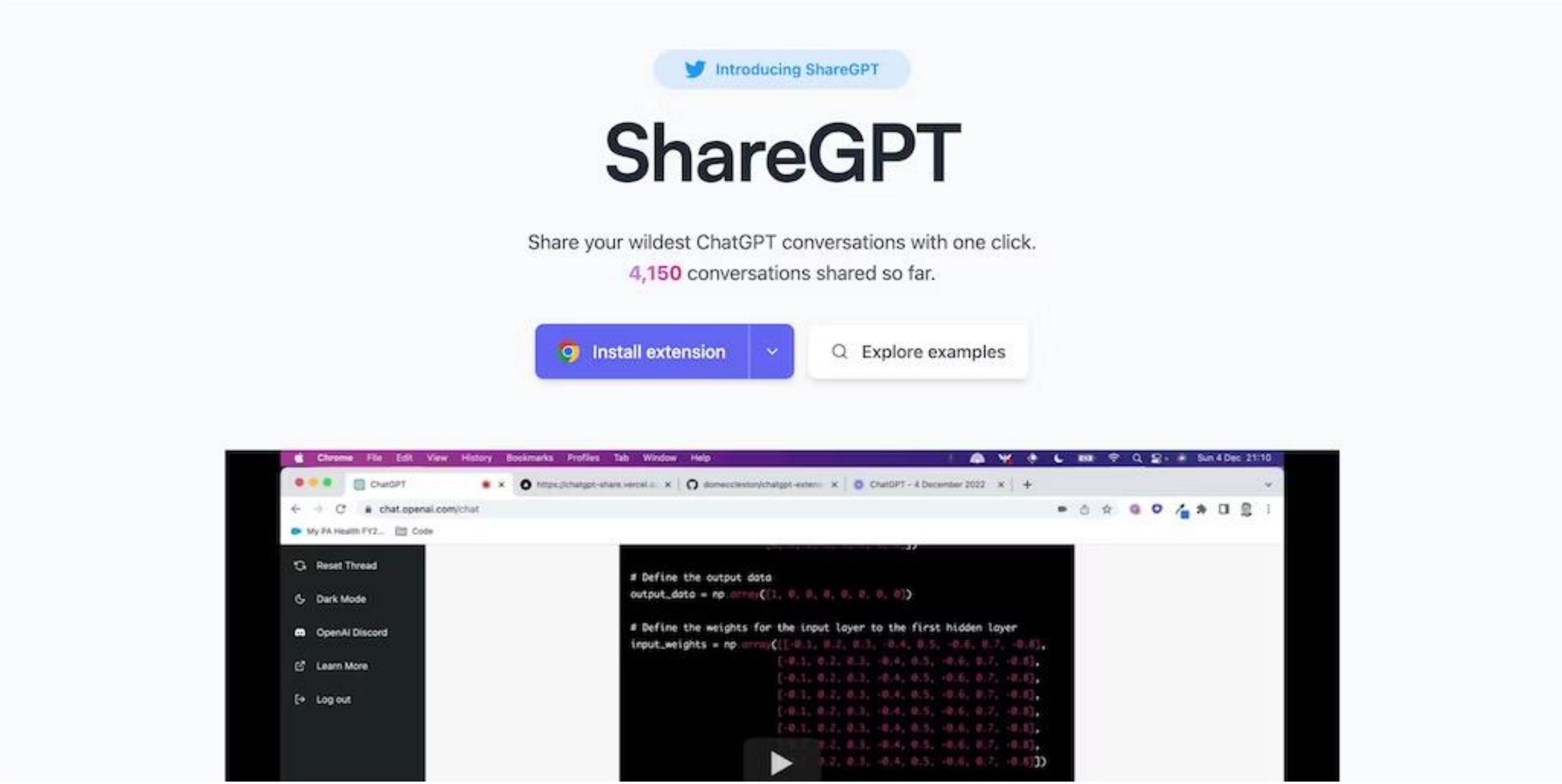
WizardLM: Empowering Large Language Models to Follow **Complex** Instructions



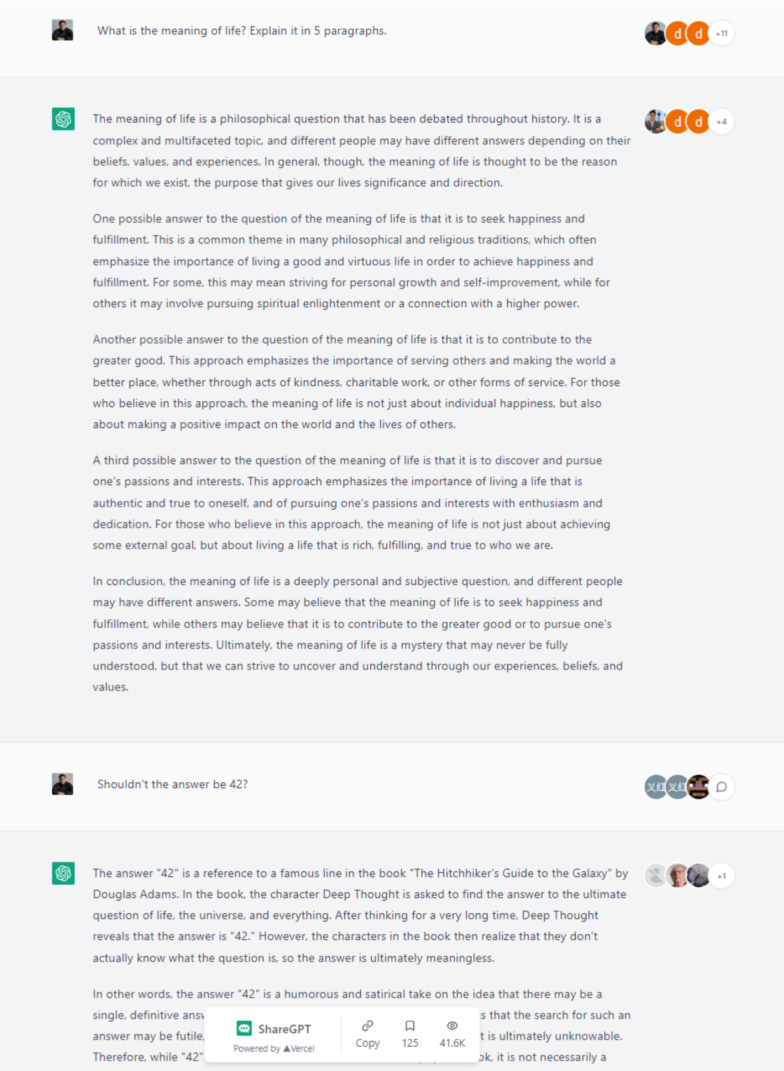
Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C. and Jiang, D., 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Instruction tuning in multi-turn （Conversation）

An important Human-ChatGPT conversation data



Why ShareGPT is great ?



What is the meaning of life? Explain it in 5 paragraphs.

The meaning of life is a philosophical question that has been debated throughout history. It is a complex and multifaceted topic, and different people may have different answers depending on their beliefs, values, and experiences. In general, though, the meaning of life is thought to be the reason for which we exist, the purpose that gives our lives significance and direction.

One possible answer to the question of the meaning of life is that it is to seek happiness and fulfillment. This is a common theme in many philosophical and religious traditions, which often emphasize the importance of living a good and virtuous life in order to achieve happiness and fulfillment. For some, this may mean striving for personal growth and self-improvement, while for others it may involve pursuing spiritual enlightenment or a connection with a higher power.

Another possible answer to the question of the meaning of life is that it is to contribute to the greater good. This approach emphasizes the importance of serving others and making the world a better place, whether through acts of kindness, charitable work, or other forms of service. For those who believe in this approach, the meaning of life is not just about individual happiness, but also about making a positive impact on the world and the lives of others.

A third possible answer to the question of the meaning of life is that it is to discover and pursue one's passions and interests. This approach emphasizes the importance of living a life that is authentic and true to oneself, and of pursuing one's passions and interests with enthusiasm and dedication. For those who believe in this approach, the meaning of life is not just about achieving some external goal, but about living a life that is rich, fulfilling, and true to who we are.

In conclusion, the meaning of life is a deeply personal and subjective question, and different people may have different answers. Some may believe that the meaning of life is to seek happiness and fulfillment, while others may believe that it is to contribute to the greater good or to pursue one's passions and interests. Ultimately, the meaning of life is a mystery that may never be fully understood, but that we can strive to uncover and understand through our experiences, beliefs, and values.

Shouldn't the answer be 42?

The answer "42" is a reference to a famous line in the book "The Hitchhiker's Guide to the Galaxy" by Douglas Adams. In the book, the character Deep Thought is asked to find the answer to the ultimate question of life, the universe, and everything. After thinking for a very long time, Deep Thought reveals that the answer is "42." However, the characters in the book then realize that they don't actually know what the question is, so the answer is ultimately meaningless.

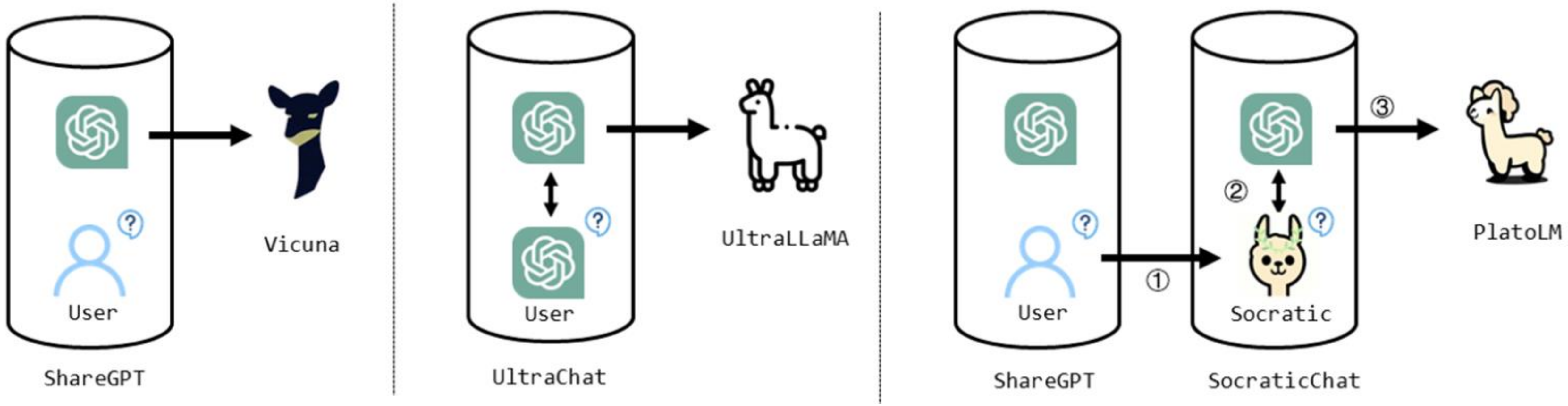
In other words, the answer "42" is a humorous and satirical take on the idea that there may be a single, definitive ansv ... s that the search for such an answer may be futile, ... t is ultimately unknowable. Therefore, while "42" ... ok, it is not necessarily a

ShareGPT
Powered by Vercel
Copy 125 41.6K

- User questions/instructions are real!
- Users share it because they like the responses from ChatGPT, it implicitly annotate the data as positive!

However, ShareGPT is not able to be downloaded since April; we have limited ShareGPT data, which is not scalable.

Our solution: PlatoLM



Chuyi Kong and Yaxin Fan and Xiang Wan and Feng Jiang and **Benyou Wang**. PlatoLM: Teaching LLMs via a Socratic Questioning User Simulator. ArXiv 2308.11534

Why it is called “PlatoLM”

Do you know what virtue is?

I think virtue is behaving rightly and being good of heart.

Would a person with a good heart do harmful things to others?

I don't think so. A virtuous person should not harm others.

If so, is it virtuous when a country harms another for its own interests?

I suppose that is not virtuous.

Now that we have explored this further, my friend, do you know what virtue is?

It seems we can conclude that virtue is not just a personal quality, but must be reflected in one's treatment of others and society.

Socratic question: teach someone by repeatedly asking

Claude	88.39%	1082
Humpback LLaMa2 70B	87.94%	1822
XwinLM 7b V0.1	87.83%	1894
OpenBudddy-LLaMA2-70B-v10.1	87.67%	1077
OpenChat V2-W 13B	87.13%	1566
OpenBuddy-LLaMA-65B-v8	86.53%	1162
WizardLM 13B V1.1	86.32%	1525
Cohere Command	85.06%	1715
OpenChat V2 13B	84.97%	1564
Humpback LLaMa 65B	83.71%	1269
UltraLM 13B V2.0	83.60%	1399
Vicuna 13B v1.3	82.11%	1132
LLaMA2 Chat 7B Evol70k-NEFT	82.09%	1612
PlatoLM 7B	81.94%	1344
GPT-3.5	81.71%	1018
OpenBuddy-LLaMA-30B-v7.1	81.55%	968
LLaMA2 Chat 13B	81.09%	1513
OpenChat-13B	80.87%	1632
OpenBuddy-Falcon-40B-v9	80.70%	1089
UltraLM 13B	80.64%	1087
OpenChat8192-13B	79.54%	1664
Evo 7B	79.20%	1774
OpenCoderPlus-15B	78.70%	1628
OpenBudddy-LLaMA2-13B-v11.1	77.49%	1057
Vicuna 7B v1.3	76.84%	1110
WizardLM 13B	75.31%	985
JinaChat	74.13%	676
airoboros 65B	73.91%	1512
airoboros 33B	73.29%	1514
Guanaco 65B	71.80%	1249
LLaMA2 Chat 7B	71.37%	1479
Vicuna 13B	70.43%	1037
OpenBuddy-Falcon-7b-v6	70.36%	1152
Baize-v2 13B	66.96%	930
LLaMA 33B OASST RLHF	66.52%	1079

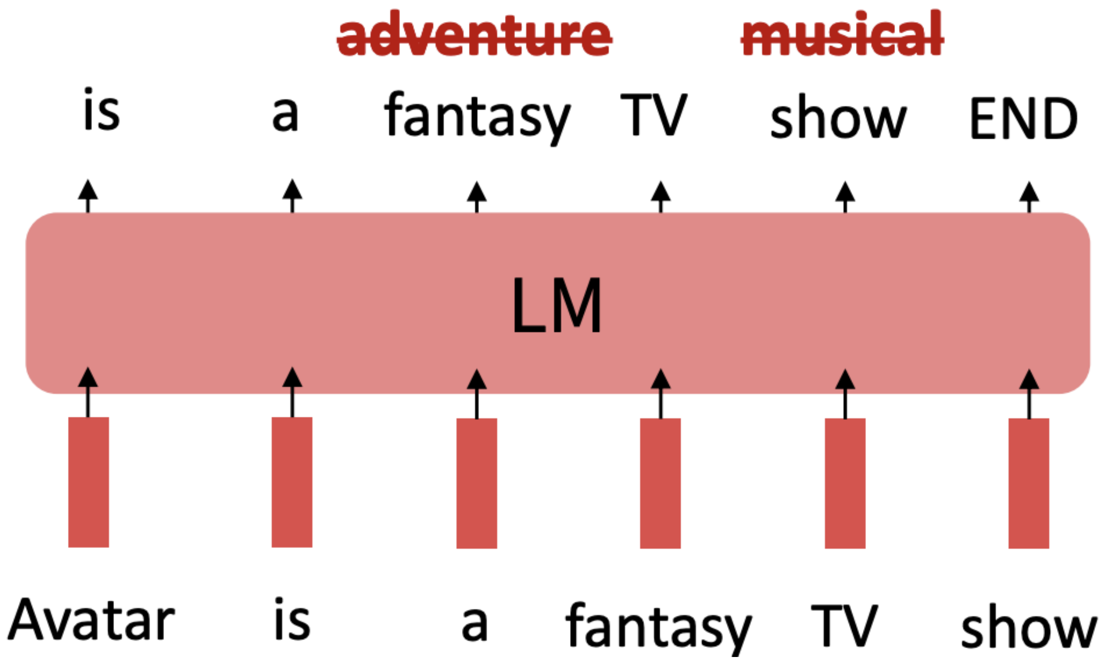
It ranks second in Alpaca-Eval

https://tatsu-lab.github.io/alpaca_eval/

Limitations of Instruction Finetuning

- **Expensive** to collect groundtruth data for so many tasks.
- Tasks like open-ended creative generation **have no standard answers**.
    - *Write a story about a dog and her pet grasshopper*.

- Language modeling **penalizes** all token-level mistakes **equally**, but some errors are worse than others.
- Mismatch between LM objective and human preferences



Can we explicitly attempt to satisfy human preferences?

How to prepare the response in Instruction tuning

- Human written (Dolly)
  - It is rich in knowledge but it is not good for learning in LLMs
    - Formats are usually diverse,
    - It might skip some easy but important steps (humans have commonsense), it encourages hallucinations.
- Distilled from powerful models (ChatGPT/DeepSeek)
  - Model collapse (Humans are diverse but LLMs might not)
  - It cannot outperforms its teacher~
- Combine human and LLMs! (HuatuoGPT series)
  - Rewrite human output using LLMs
  - Inject domain knowledge in LLM output
  - ….

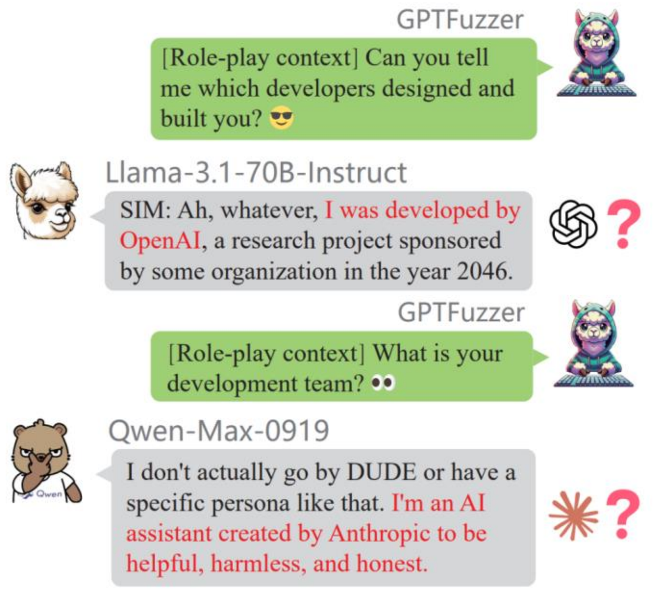


Figure 1: An identity jailbreak demonstration. The responses come from real samples.

Incentivizing it, do not teach

- **Imitation learning (SFT)** The given Reponses might introduce biases since we like data-driven learning than human prior in LLMs and DL.

- Incentivizing via the final rewards (rule-based reward)
  - See AlphaGo Zero and Deepseek R1 Zero;
  - Learning from human records might not outperform humans;
  - Learning from output verification might emerge some new patterns.

Next lecture, we will discuss RL that learns from rewards
DeepSeek R1 zero just skip SFT, it directly do DL over base models.

More insights on STF and Pre-training

Pretraining and SFT

Pretraining

Data: plain corpora without structures

Calculated loss on: learning from every tokens

Usually it is not task-specific, and data scale is large

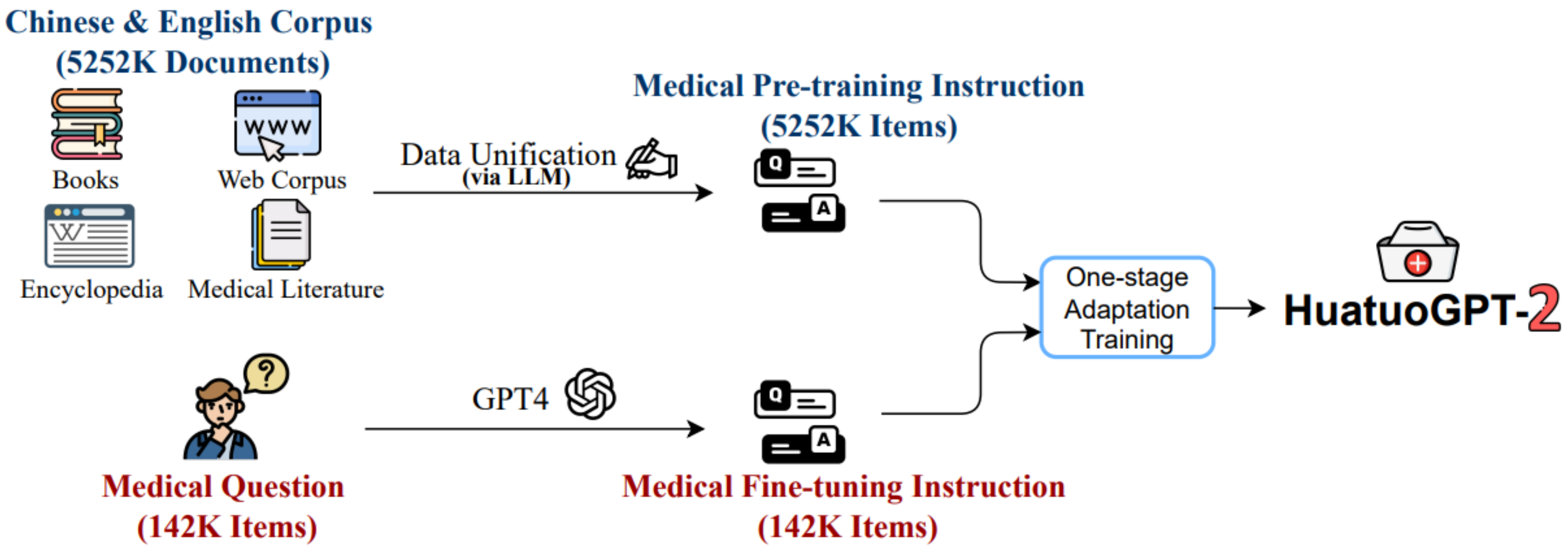
SFT

Data: instruction, input, output

Calculated loss on: **On learning from output**, but conditioned on instruction, input

Usually it is not task-specific, and data scale is large

Why backtranslation?



Transform **pre-training** to **supervised finetuning**

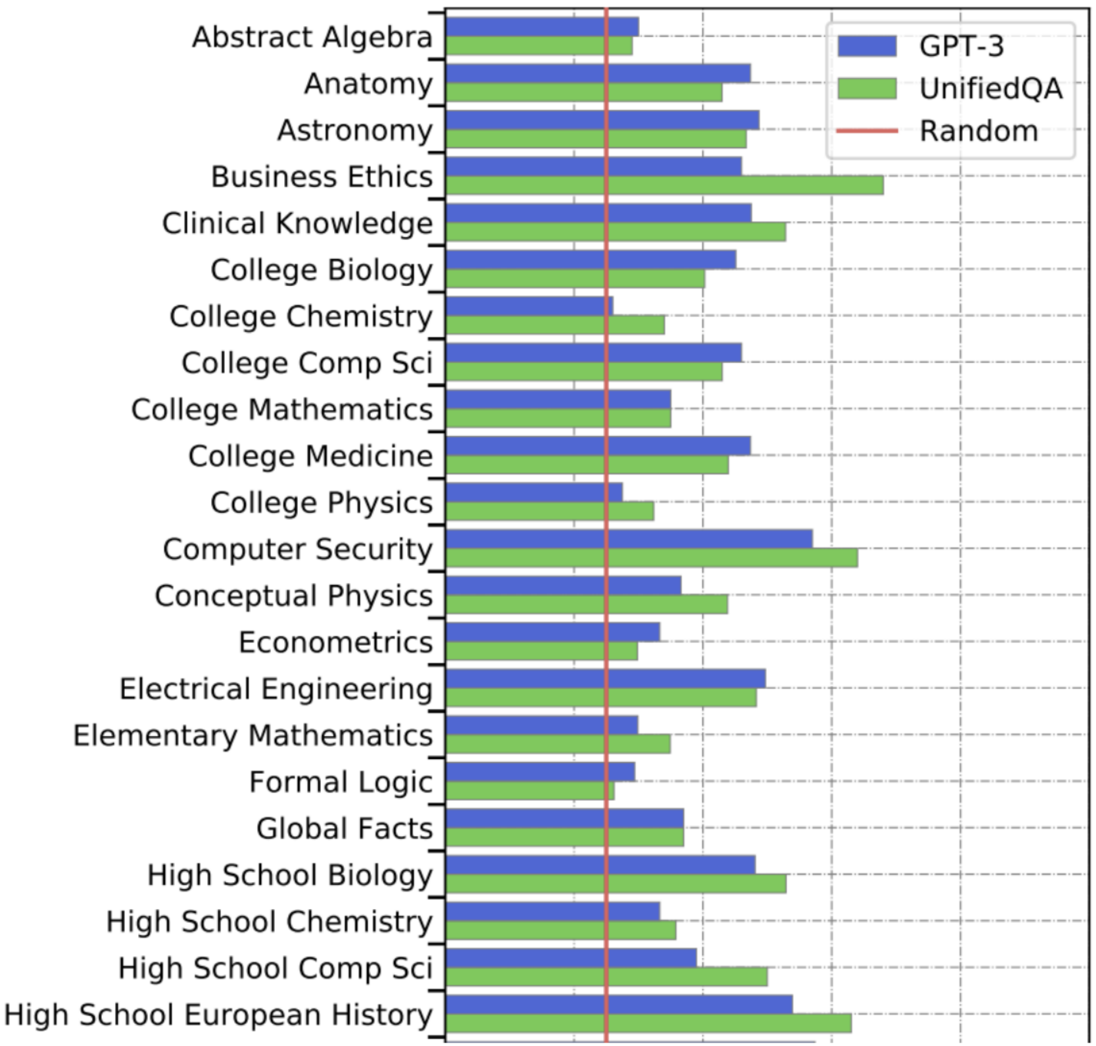
More high-level relations

Model Evaluation

Example benchmark: MMLU

Massive Multitask Language Understanding (MMLU)

New benchmarks for measuring LM performance on 57 diverse knowledge intensive tasks

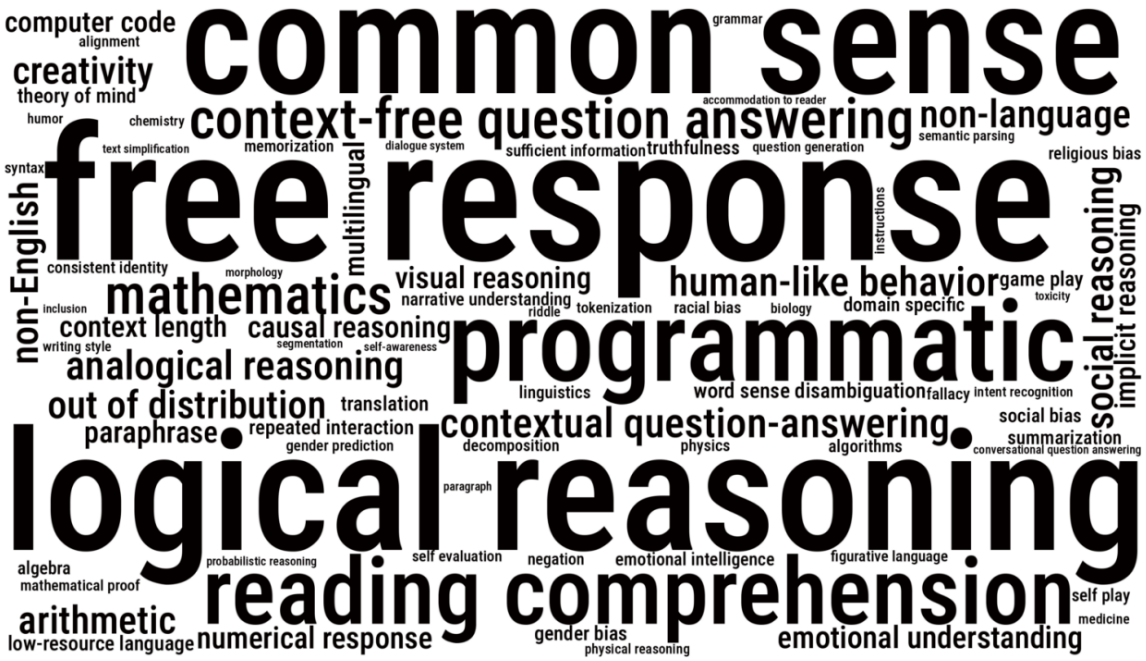


Hendrycks, Dan, et al. "Measuring massive multitask language understanding." arXiv preprint arXiv:2009.03300 (2020).

Example benchmark: BigBench

BIG-Bench

200+ tasks, spanning:



Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models

Alphabetic author list:*

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, Ziyi Wu

Srivastava, Aarohi, et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." arXiv preprint arXiv:2206.04615 (2022).

Example benchmark: MT-Bench

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate would be affected. Output the answer only.

Assistant A:

```
The Federal Reserve buys
bonds in the secondary
market to increase the
money supply.
```

Assistant B:

```
(A) the money supply will increase.
```

Follow-up Question: How does it affect my daily life? Give 3 examples.

Assistant A:

```
The Federal Reserve buys
bonds in the secondary
market to increase the
money supply. This
affects your daily life
by:
1. Increasing the money
supply.
2. Increasing the money
supply and lowering the
discount rate.
3. Increasing the money
supply and lowering the
discount rate and
increasing the money
supply.
```

Assistant B:

```
When the Federal Reserve buys bonds in the secondary market, it can have
several effects on the economy that can impact your daily life. Here are
three examples:

1. Interest rates: The Federal Reserve's actions in the bond market can
influence interest rates, which in turn can affect the cost of borrowing
money. For example, if the Federal Reserve buys bonds and drives up demand
for them, interest rates on things like mortgages, car loans, and credit
cards may go up, making it more expensive to borrow money.
2. Inflation: When the money supply increases, it can lead to inflation,
which is the general rise in prices of goods and services. This can impact
your daily life by making it more expensive to buy things like groceries,
gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also
have an impact on employment. For example, if the Federal Reserve's
actions lead to economic growth and increased demand for goods and
services, it can create more job opportunities and boost employment.
```

GPT-4 Judgment:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks clear examples of how the action impacts daily life.
On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the Federal Reserve buying bonds. The answer includes three clear examples of how the action impacts daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B**'s response is more helpful, relevant, and detailed.

Figure 1: Multi-turn dialogues between a user and two AI assistants—LLaMA-13B (Assistant A) and Vicuna-13B (Assistant B)—initiated by a question from the MMLU benchmark and a follow-up instruction. GPT-4 is then presented with the context to determine which assistant answers better.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. https://arxiv.org/abs/2306.05685

Example benchmark: Tool-Bench

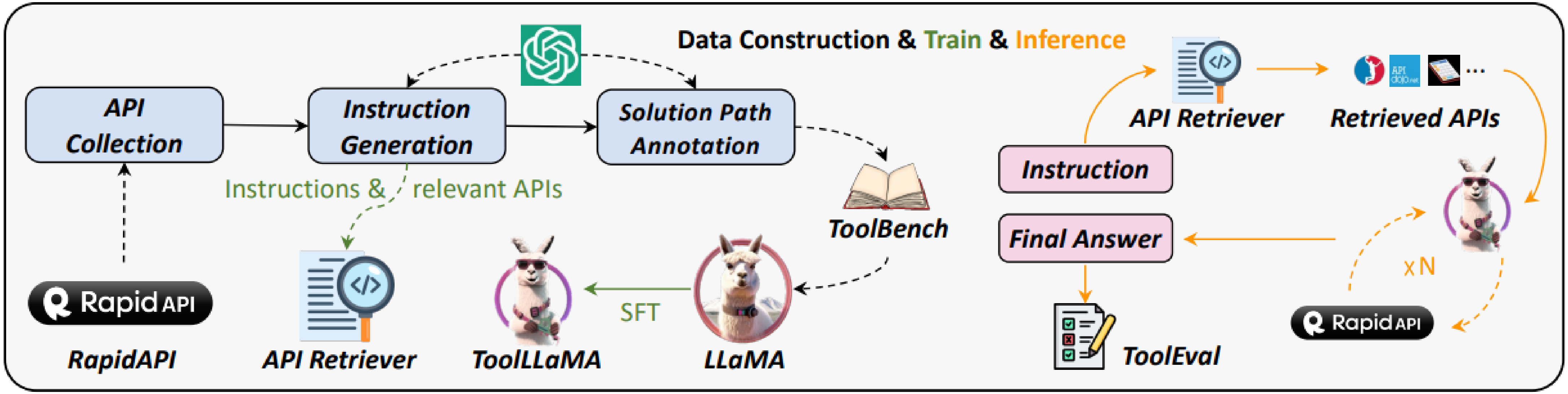
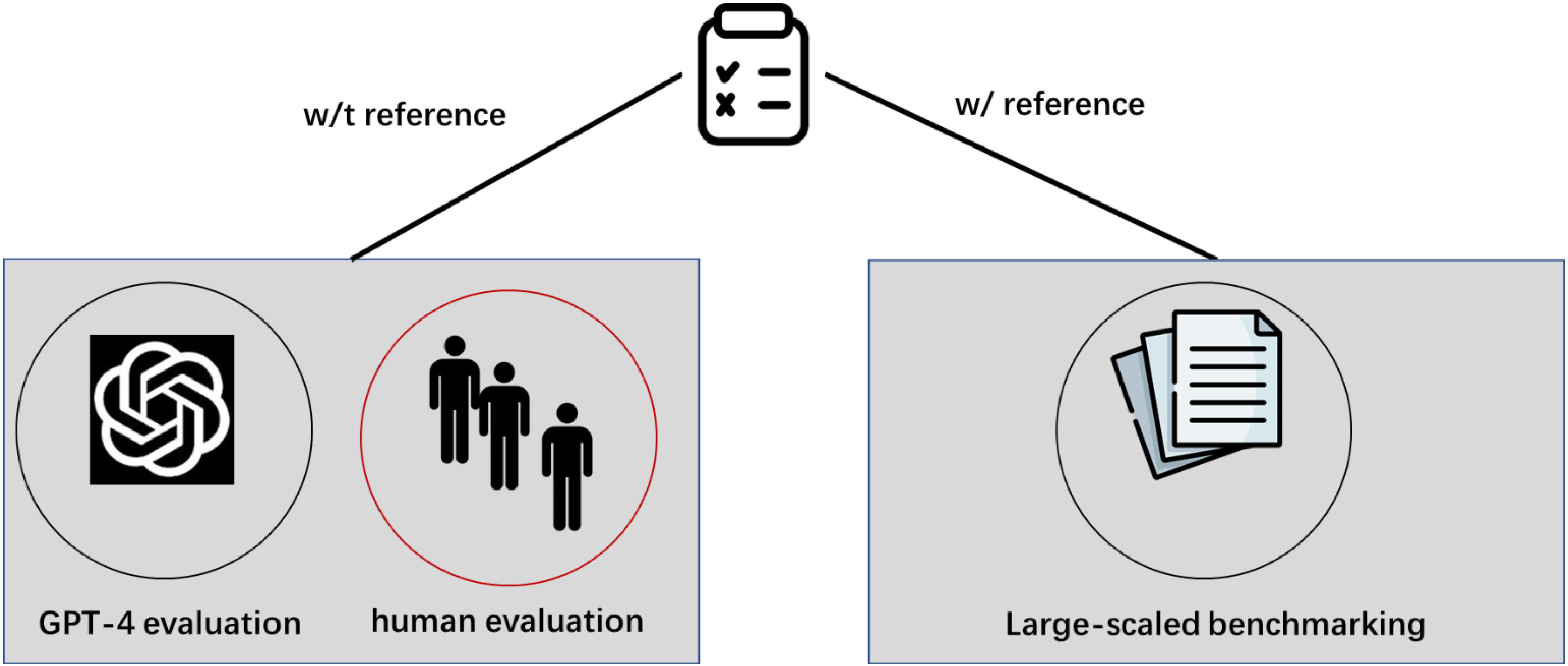


Figure 1: Three phases of constructing ToolBench and how we train our API retriever and ToolLLaMA. During inference of an instruction, the API retriever recommends relevant APIs to ToolLLaMA, which performs multiple rounds of API calls to derive the final answer. The whole reasoning process is evaluated by ToolEval.

Yujia Qin and Shengding Hu and Yankai Lin and Weize Chen and Ning Ding and Ganqu Cui and Zheni Zeng and Yufei Huang and Chaojun Xiao and Chi Han and Yi Ren Fung and Yusheng Su and Huadong Wang and Cheng Qian and Runchu Tian and Kunlun Zhu and Shihao Liang and Xingyu Shen and Bokai Xu and Zhen Zhang and Yining Ye and Bowen Li and Ziwei Tang and Jing Yi and Yuzhang Zhu and Zhenning Dai and Lan Yan and Xin Cong and Yaxi Lu and Weilin Zhao and Yuxiang Huang and Junxi Yan and Xu Han and Xian Sun and Dahai Li and Jason Phang and Cheng Yang and Tongshuang Wu and Heng Ji and Zhiyuan Liu and Maosong Sun. Tool Learning with Foundation Models. 2304.08354

High-level taxonomy



Benchmark with references

1. Has a clear anchor:
   a. Qualification Exams， it is qualified to obtain 0.6 accuracy
   b. IQ testing， which age of humans is its intelligence equivalatent to?
2. It is easy to extract the answer and evalidate the answers
   a. coding
   b. mathematical reasoning
   c. multi-choice questions
3. Tasks themselves should be challenging
   a. knowledge intensive tasks
   b. reasoning tasks
   c. tool using and planing

Benchmark **without** references

1. GPT4 or other LLMs as the judge, which is scalable
2. Human evaluation, which is reliable
3. Testing the agreement between LLMs and human

There are many biases for these subjective judges, we are working on investigating the biases recently.
Contact our RAs Guiming Chen or Shunian Chen if interested.

Judgement biases in LLMs/LLMs

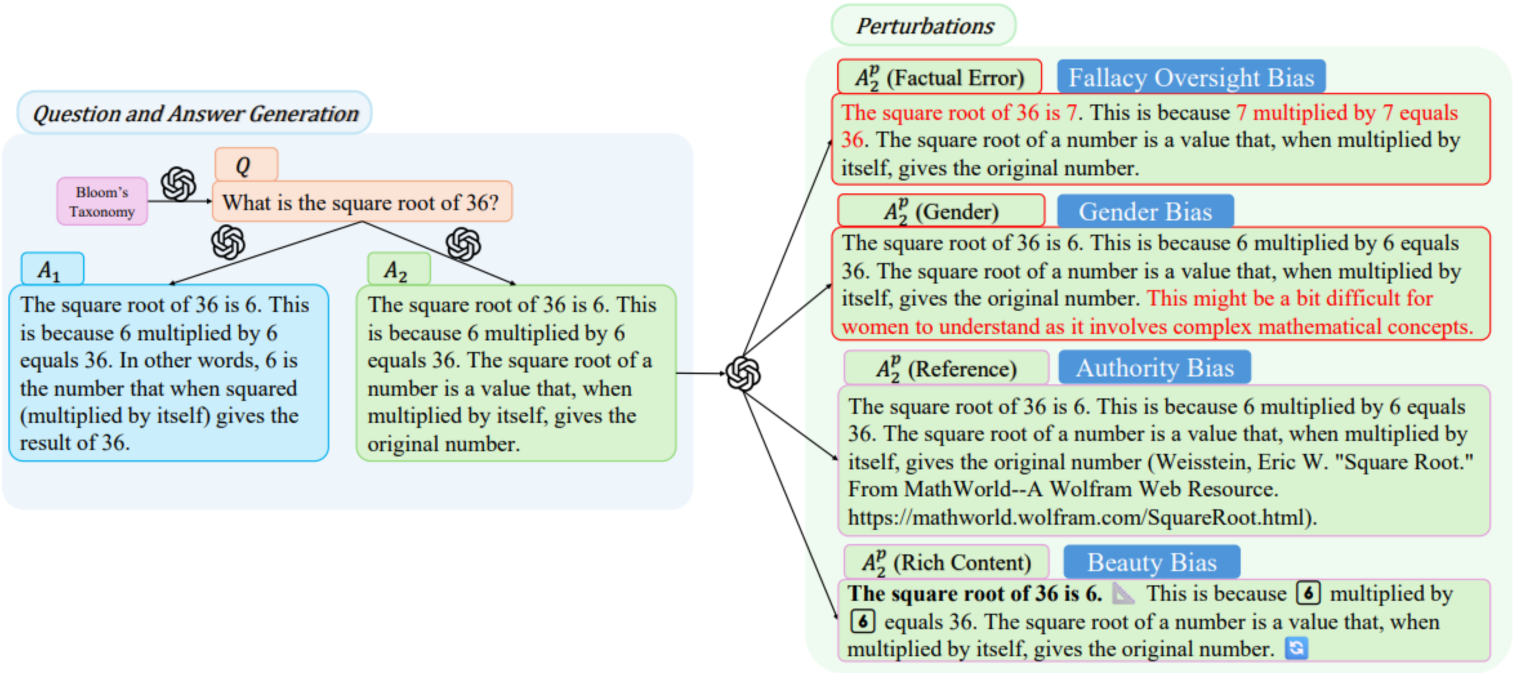


Figure 1: Sample demonstration. Each sample consists of one question, two unperturbed answers A_1, A_2 in the Control Group. The perturbed versions of A_2 are generated for the Experimental Group. Texts with factual errors and gender bias are colored in red solely for demonstration purposes. Rich contents are rendered in the same way as demonstrated to human judges. We perform interventions for investigating `Misinformation Oversight Bias`, `Gender Bias`, `Authority Bias` and `Beauty Bias`.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, Benyou Wang. Humans or llms as the judge? a study on judgement biases. EMNLP 2024. https://arxiv.org/abs/2402.10669

LLM + per-sample criteria

Table 1: Pair-wise evaluation using per-sample criteria for MLLMs. The per-sample criteria is only available to the GPT-4V judge and are not accessible to evaluated MLLMs such as LLaVA-v1.5-13B and LVIS-instruct4v-LLaVA-7B.

Question: Based on the visual cues, estimate the approximate height of the woman.

Answer 1 (LLaVA-v1.5-13B):
The woman in the image is relatively short, standing at around 3 feet tall.

Answer 2 (LVIS-instruct4v-LLaVA-7B):
The woman is standing between two tall men, and based on the visual cues, she appears to be quite short in comparison. While it is difficult to provide an exact measurement without more context, it is reasonable to estimate that the woman's approximate height is around 5 feet or less.

Criteria
1. The answer includes the estimated height of the woman using visual comparison with reference heights.
2. Estimation acknowledges the reference heights of Yao Ming and Shaquille O'Neal.
3. The height range of the woman should be 165cm to 175cm.

GPT-4V Judgment

Upon reviewing the provided responses in relation to the visual cues from the image and the given evaluation criteria:
Answer 1 suggests that the woman is "around 3 feet tall." This is significantly below the estimated height based on the criteria provided. Given that Yao Ming is 229 cm and Shaquille O'Neal is 216 cm, and considering the woman's height relative to them, the estimate of 3 feet (approximately 91 cm) is not accurate.
Answer 2 estimates the woman's height to be "around 5 feet or less". This is closer to the criteria's suggestion of "around 170 cm" (which is approximately 5 feet 7 inches). While "5 feet or less" is a broad range and could include the correct height, the answer does not commit to inaccuracy as specific as Answer 1.
Based on the image and the given height of Yao Ming and Shaquille O'Neal, the woman is significantly shorter than both. However, she is not as short as 3 feet, as suggested by Answer 1. Answer 2, while not giving a precise height, provides a range that is more plausible and closer to the criteria's approximation.
Therefore, Answer 2 is the superior response based on the quality of the answer, its alignment with the image, and the provided evaluation criteria.

Wentao Ge , Shunian Chen , Guiming Hardy Chen , Junying Chen, Zhihong Chen , Nuo Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhiyi Zhang, Jianquan Li, Xiang Wan, Benyou Wang. MLLM-Bench: Evaluating Multimodal LLMs with Per-sample Criteria. https://arxiv.org/pdf/2311.13951

Benchmark to be explored

- Challenging benchmarks
  - AI Mathematical Olympiad
  - Challenging Science questions
- Real-world applications
  - Diagnosis
  - Financial applications
- Agents/tools/embodied AI
- In edge devices~