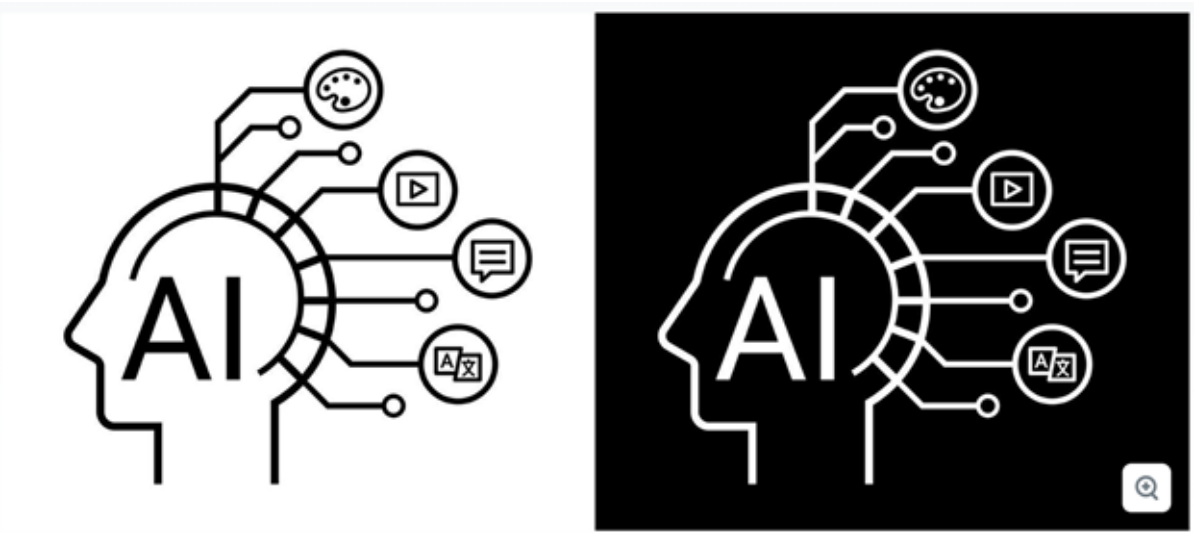
AIMS 5740
Generative Artificial Intelligence

(2026 Term 2)



Computer Science & Engineering
The Chinese University of Hong Kong

Announcement

- Our first assignment will be released on Feb 5. We will have a tutorial on Feb 4 for that.

- Our course project will be released on late Feb. A team is suggested to have 3-4 people (maximum 4).

Generative Models

- Learning to generate data

Samples from a Data Distribution

Train

Neural Network

Sample

Generative Models

Given training data, generate new samples from same distribution

Training data ~ $p_{data}(x)$

Generated samples ~ $p_{model}(x)$

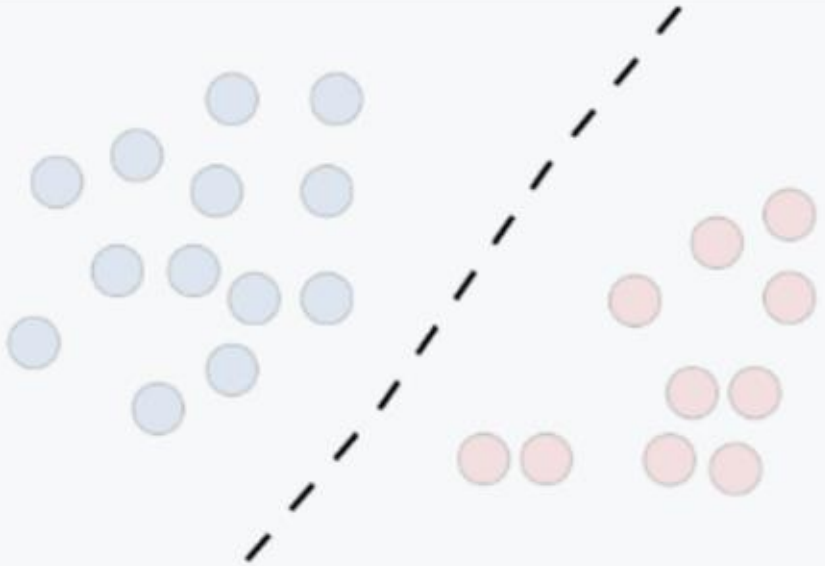
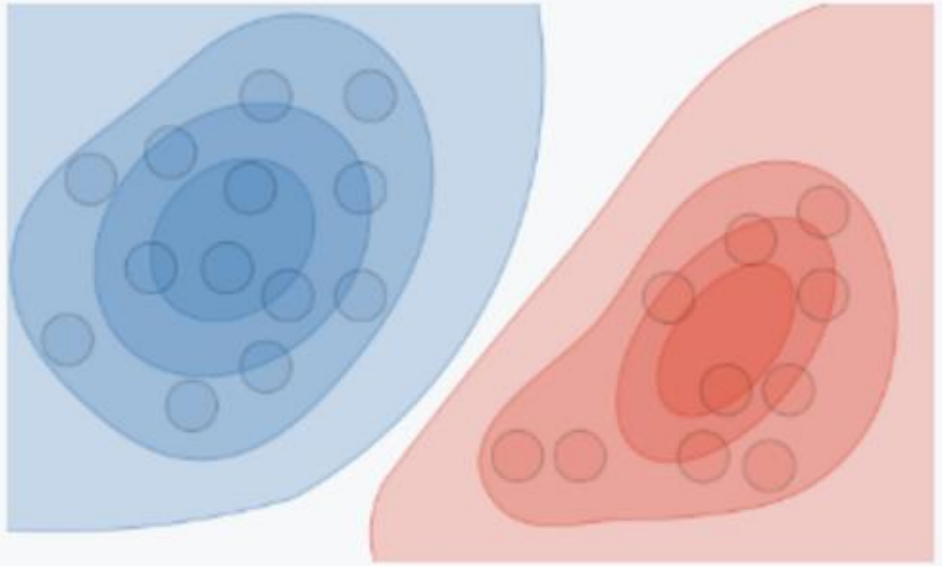
Want to learn $p_{model}(x)$ similar to $p_{data}(x)$

- A core problem in unsupervised learning
- Different flavors:

Explicit density estimation: explicitly define and solve for P_{model}

Implicit density estimation: learn model that can sample from P_{model}

Generative vs. Discriminative

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		

P(x|y) or P(y|x) ?

- Learning to generate data

Samples from a Data Distribution

Train

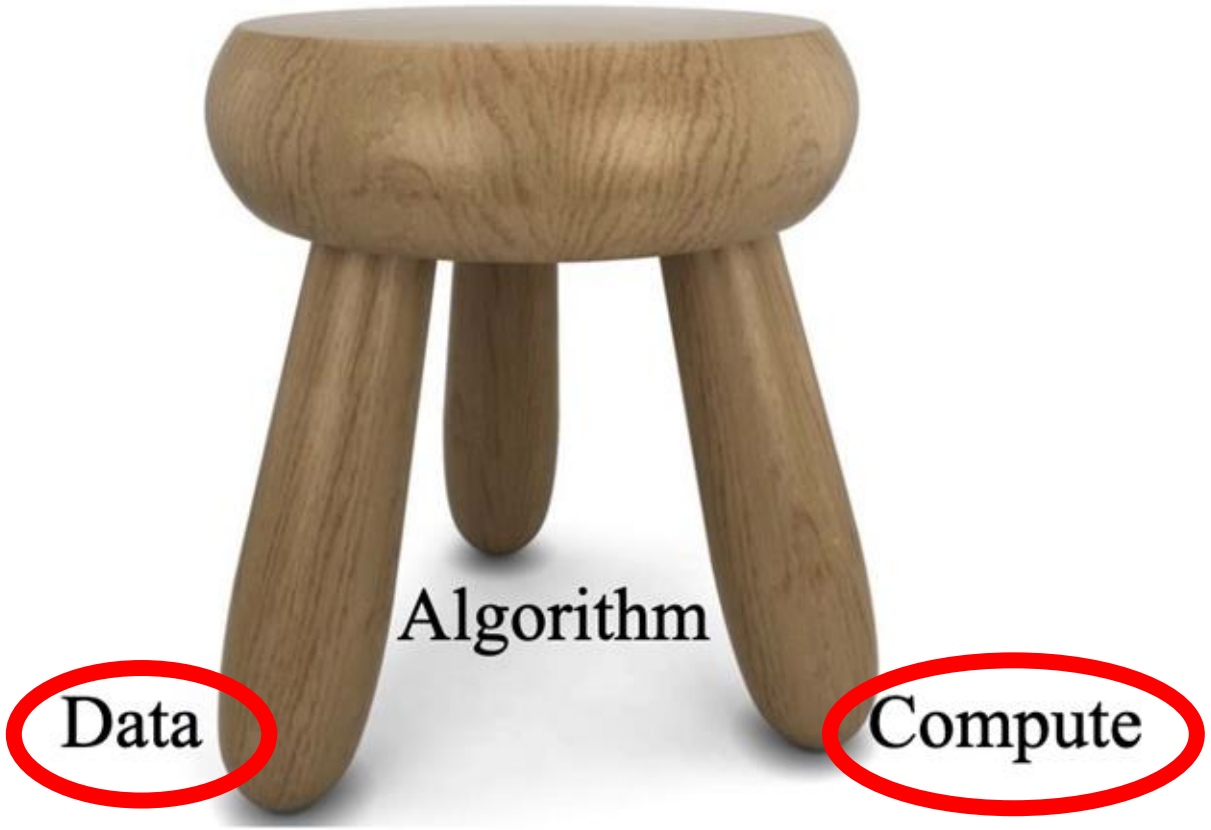
Neural Network

Cat?

Sample

Generative AI

- Data, algorithm and compute are three key components in the current generative AI applications.



GPT

- Transformer, Autoregressive modeling, 117 million parameters

- Pre-training data: BooksCorpus and 1B word Benchmark
- Fine-tuning data:

Table 1: A list of the different tasks and datasets used in our experiments.

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.9	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	88.1	91.3	45.4	80.0	82.3	56.0	75.2
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.1	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**91.1**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**81.9**

GPT 2

- Pretraining data: a massive, diverse collection of over 40GB of text scraped from the web, specifically 8 million unique web pages from a dataset called WebText.

	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	**21.8**
117M	**35.13**	45.99	**87.65**	**83.4**	**29.41**	65.85	1.16	1.17	37.50	75.20
345M	**15.60**	55.48	**92.35**	**87.1**	**22.76**	47.33	1.01	**1.06**	26.37	55.72
762M	**10.87**	**60.12**	**93.45**	**88.0**	**19.93**	**40.31**	**0.97**	**1.02**	22.05	44.575
1542M	**8.63**	**63.24**	**93.30**	**89.05**	**18.34**	**35.76**	**0.93**	**0.98**	**17.48**	42.16

Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

GPT 3

- GPT-3 has 96 layers with each layer having 96 attention heads. The size of word embeddings was increased to 12888 for GPT-3 from 1600 for GPT-2. The context window size was increased from 1024 for GPT-2 to 2048 tokens for GPT-3.

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}

GPT 3

Feature	GPT-2	GPT-3
Parameters	1.5 Billion	175 Billion
Training Data	40 GB of internet text	Hundreds of GBs of internet text

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:
We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she started doing these crazy farduddles.
A "yalubalu" is a type of vegetable that looks like a big pumpkin. An example of a sentence that uses the word yalubalu is:
I was on a trip to Africa and I tried this yalubalu vegetable that was grown in a garden there. It was delicious.

A "Burringo" is a car with very fast acceleration. An example of a sentence that uses the word Burringo is:
In our garage we have a Burringo that my father drives to work every day.

A "Gigamuru" is a type of Japanese musical instrument. An example of a sentence that uses the word Gigamuru is:
I have a Gigamuru that my uncle gave me as a gift. I love to play it at home.

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Context →	Q: What is 95 times 45? A:
Target Completion →	4275

Figure G.45: Formatted dataset example for Arithmetic 2Dx

Context →	Q: What is 509 minus 488? A:
Target Completion →	21

Figure G.46: Formatted dataset example for Arithmetic 3D-

Context →	Q: What is 556 plus 497? A:
Target Completion →	1053

Figure G.47: Formatted dataset example for Arithmetic 3D+

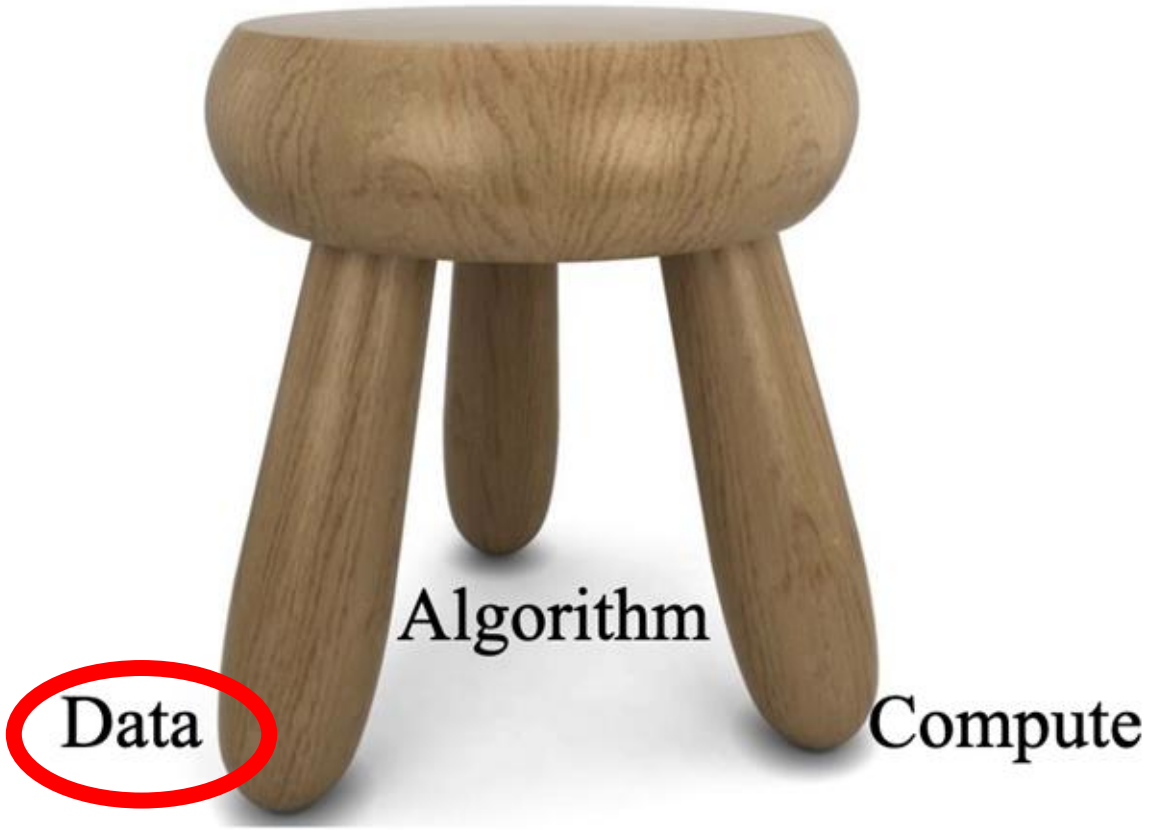
Context →	Q: What is 6209 minus 3365? A:
Target Completion →	2844

Figure G.48: Formatted dataset example for Arithmetic 4D-

Context →	Q: What is 65360 plus 16204? A:
Target Completion →	81564

Generative AI

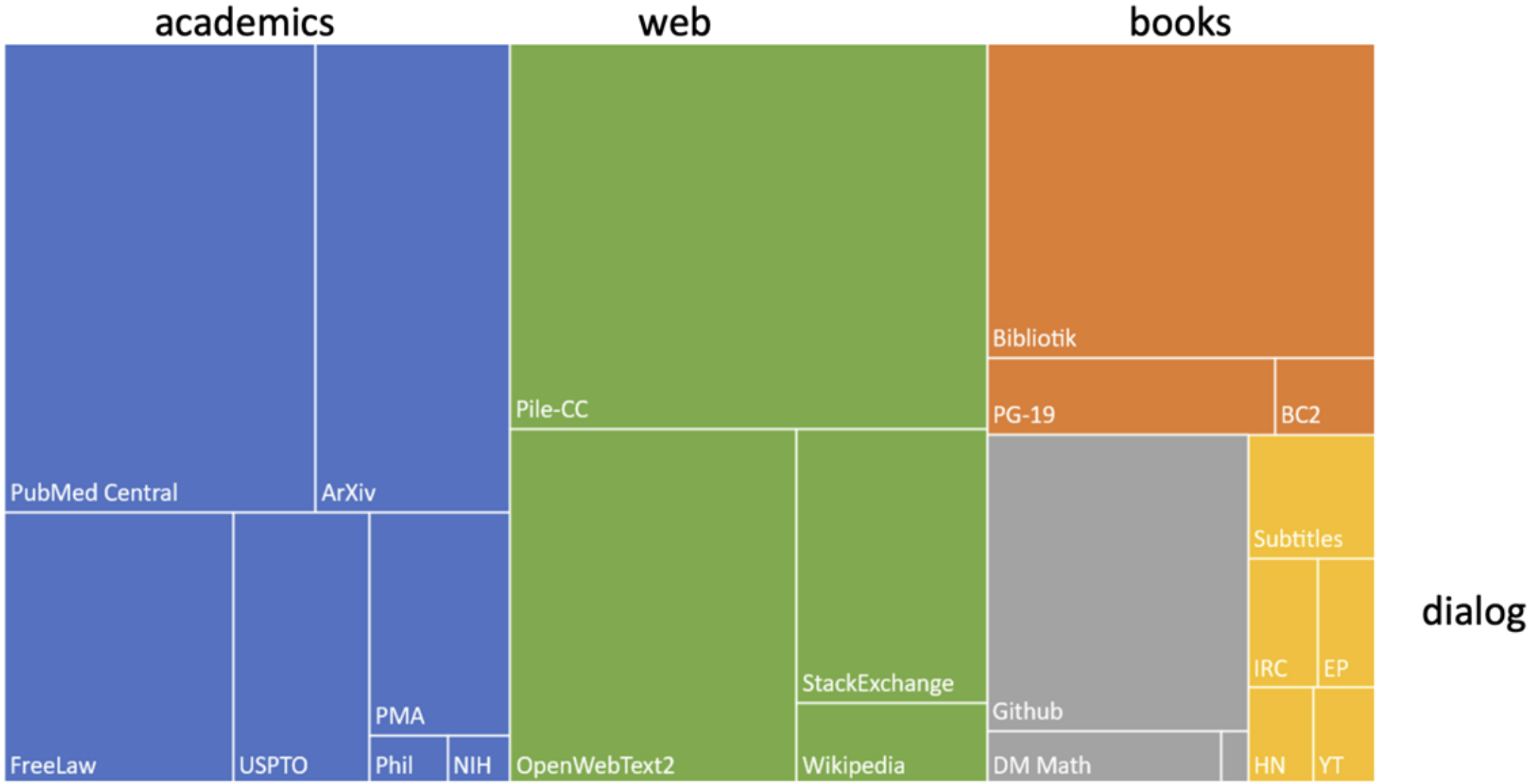
- Data is one of the most key components for GenAI.



Pre-training Data for LLMs

- **Common crawl**, snapshots of the entire web produced by the non-profit Common Crawl with billions of pages
- **Colossal Clean Crawled Corpus (C4**; Raffel et al. 2020), 156 billion tokens of English, filtered What's in it? Mostly patent text documents, Wikipedia, and news sites
- Professional data: **BookCorpus,** Exam **(**Stem, Liberal Arts), ProQuest, arXiv

The Pile: a Pre-training Corpus



Issues of Web Data

Quality is subjective

- Many LLMs attempt to match Wikipedia, books, particular websites
- Need to remove boilerplate, adult content • Deduplication at many levels (URLs, documents, even lines)

Safety also subjective

- Toxicity detection is important, although that has mixed results
- Can mistakenly flag data written in dialects like African American English

Issues of Web Data

Organization that crawls web and releases snapshots

- Still orders of magnitude below Google
- But really big!

Crawl date	Size in TiB	Billions of pages	Comments
June 2023	390	3.1	Crawl conducted from May 27 to June 11, 2023
April 2023	400	3.1	Crawl conducted from March 20 to April 2, 2023
February 2023	400	3.15	Crawl conducted from January 26 to February 9, 2023
December 2022	420	3.35	Crawl conducted from November 26 to December 10, 2022
October 2022	380	3.15	Crawl conducted in September and October 2022

https://commoncrawl.org/

Web Data

Preview Code Blame 163 lines (110 loc) · 17.8 KB Raw

language Python license Apache-2.0 pypi v1.4.5 Docker v1.4.5 Docker OSS latest Coverage 79%

DataModality Text,Image,Audio,Video Usage Cleaning,Synthesis,Analysis downloads 27k

文档 算子池 cs.LG 1.0Paper(SIGMOD'24) cs.AI 2.0Paper(NeurIPS'25)

Data-Juicer 是一个一站式系统，面向大模型的文本及多模态数据处理。我们提供了一个基于 JupyterLab 的 Playground，您可以从浏览器中在线试用 Data-Juicer。 如果Data-Juicer对您的研发有帮助，请支持加星（自动订阅我们的新发布）、以及引用我们的工作 。

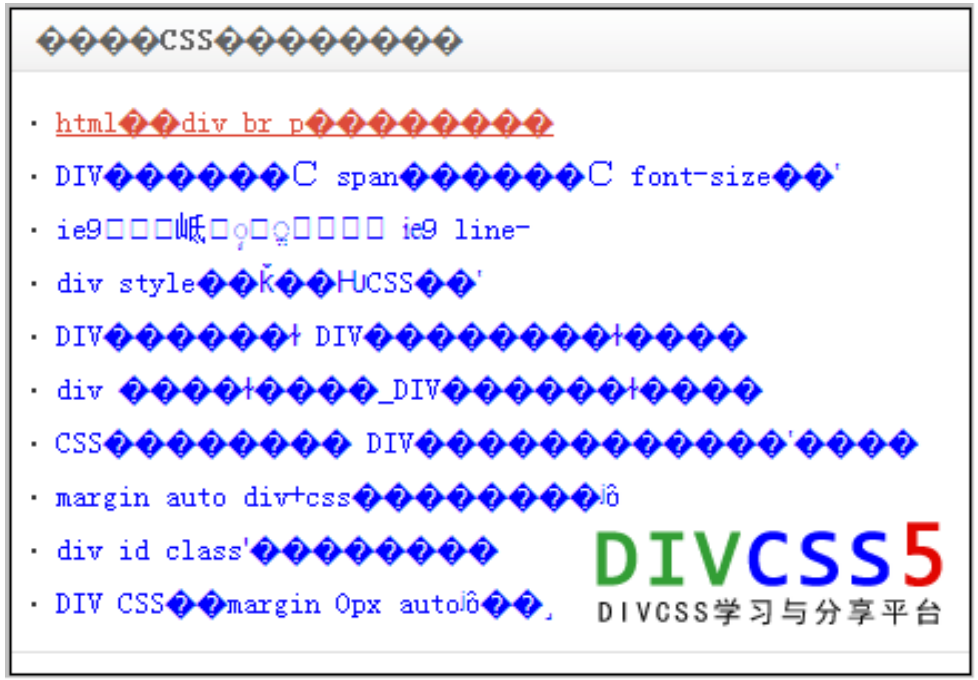
阿里云人工智能平台 PAI 已深度集成Data-Juicer到其数据处理产品中。PAI提供包含数据集管理、算力管理、模型工具链、模型开发、模型训练、模型部署、AI资产管理在内的功能模块，为用户提供高性能、高稳定、企业级的大模型工程化能力。数据处理的使用文档请参考：快速提交DataJuicer任务。

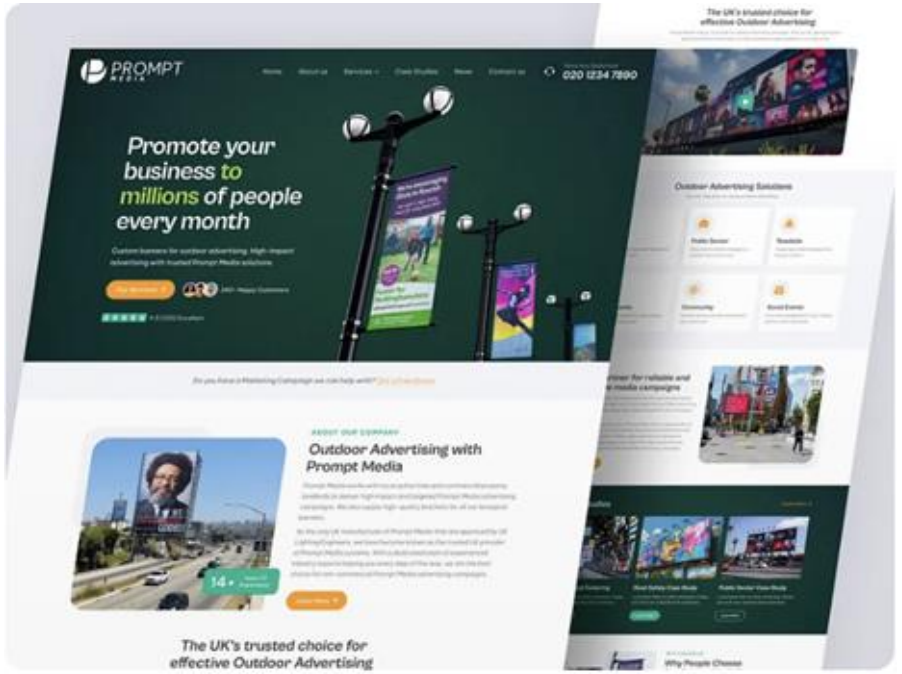
Data-Juicer正在积极更新和维护中，我们将定期强化和新增更多的功能和数据菜谱。热烈欢迎您加入我们，一起推进大模型的数据-模型协同开发和研究应用！

[Demo Video] DataJuicer-Agent:数据处理，即刻启程！

Processing Web Data (1)

- Detecting garbled, special characters
- Removing redundant and advertising content





Processing Web Data (2)

- Need to process images, tables, and hyperlink
- How?

GPT-3 相比 GPT-2，模型参数变大了很多，模型参数直接干到了1.75万亿（1750B），是 GPT-2 的1000多倍，如下图所示：

列出GPT-2 (最大版本)和GPT-3的主要参数，表格形式

以下是GPT-2 (最大版本)和GPT-3的主要参数列表：

参数	GPT-2 (最大版本)	GPT-3
模型层数	48	96/2048
隐藏层维度	2048	12288
注意力头数	16	96
参数数量	15亿	1.75万亿
输入序列长度	最大1024个token	最大2048个token
训练语料库	网络上抓取的数据和书籍	Common Crawl、维基百科、图书

之前的 GPT-1 是在子任务训练时提供少量的训练数据，GPT-2 **是处理子任务时不提供任何相关的训练样本，直接使用预训练模型在子任务上面做预测。**

GPT-3 [编辑]

31种语言

条目 讨论 汉 漢 大陆简体 阅读 编辑 查看历史 工具

维基百科，自由的百科全书

生成型预训练变换模型 3 （英语：Generative Pre-trained Transformer 3，简称 **GPT-3**）是一个自回归语言模型，目的是为了使用深度学习生成人类可以理解的自然语言[1]。GPT-3是由在旧金山的人工智能公司OpenAI训练与开发，模型设计基于谷歌开发的 Transformer 语言模型。GPT-3的神经网络包含1750亿个参数，需要700GB来存储[2]。该模型在许多任务上展示了强大的零样本和少样本的能力[3]。

OpenAI于2020年5月发表GPT-3的论文，在次月为少量公司与开发人团发布应用程序接口的测试版。微软在2020年9月22日宣布获取了GPT-3的独家授权[4]。

GPT-3被认为可写出人类无法与电脑区别的文章与字符串，GPT-3原始论文的作者们警告了GPT-3有可能对于社会的负面影响，比如利用制造假新闻的可能性。英国《卫报》即使用GPT-3生成了一个关于人工智能对人类无威胁的评论专栏[5]。李开复称卷积神经网络与GPT-3为人工智能重要的改善，两者皆是模型加海量数据的成果[6]。

生成型预训练变换模型 3
Generative Pre-trained Transformer 3（GPT-3）

原作者	OpenAI
首次发布	2020年6月11日，5年前 (beta)
当前版本	2023年2月13日，2年前
源代码库	https://github.com/openai/gpt-3
前任	GPT-2
继任	GPT-4
类型	大型语言模型 基于转换器的生成式预训练模型 基础模型
许可协议	专有
网站	openai.com/blog/openai-api

Processing Web Data (3)

- How? Captioning, OCR
- Hyperlink->crawling the new content

```
## Images and links:

Link using anchor text like [this](http://
example.com)

And use the code below to insert an image:

![](https://d1gb2mb5crnatu.cloudfront.net/
startups/i/271020-
a0984b3197c2dcc20d0aac2da035fabf-
medium_jpg.jpg?buster=1448230630)
```

Images and links:

Link using anchor text like this

And use the code below to insert an image:

```
# Formatting in markdown

Use underscores for _italics_

Double asterisks for **bold**

Bullet points like this:

* I'm a bullet point
+ I'm another one
- Me too

`<p>use a backtick on each side of a line for
code formatting</p>`

``` or surround
multiline code
with
three backticks```

~~strike me through~~

> Quote me.
>> Nest quote me.

And make a horizontal line.

***
```

FORMATTING IN MARKDOWN

Use underscores for *italics*

Double asterisks for **bold**

Bullet points like this:

- I'm a bullet point
- I'm another one
- Me too

`<p>use a backtick on each side of a line for code formatting</p>`

```
or surround
multiline code
with
three backticks
```

~~strike me through~~

> *Quote me.*
>
> > *Nest quote me.*

And make a horizontal line.

```

Processing Web Data (4)

- Structure format: markdown, title, subtitle
- Special symbols, math equations

Important equations in computing	
$\frac{df}{dt} = \lim_{(h \to 0)} \frac{f(t+h) - f(t)}{h}$	**Fundamental Theorum of Calculus** Finds optimal solutions to problems in computing and other disciplines.
$\hat{f}\ (\xi) = \sum_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx$	**Fourier Transform** Enables JPEG and MP3 compression.
$\vec{\nabla} \cdot \vec{B} = 0$ $\vec{\nabla} \cdot \vec{D} = \rho_{enc}$ $\vec{\nabla} \times \vec{E} = -\frac{\partial B}{\partial t}$ $\vec{\nabla} \times \vec{H} = \vec{J} + \frac{\partial \vec{D}}{\partial t}$	**Maxwell's Equations** Describes how modern communication uses electromagnetism and optics.
$i\hbar \frac{\partial}{\partial t} \Psi(t) = \hat{H} \Psi(t)$	**Schrödinger's equation** Describes state in quantum computing.

Data Format

• Different data formats can result in different learning speeds. Also note that they are the same data – the same equations, the same answers, just in different formats.

Data Formatting

Plain

128+367=495

Reverse

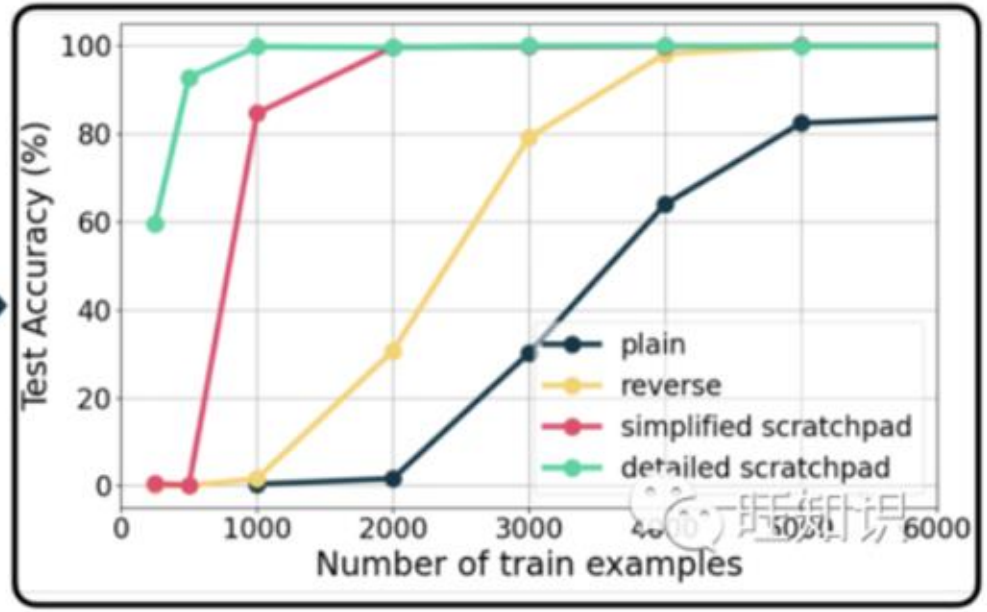
$128+367=594$

Simplified Scratchpad

Input: 128+367
Target:
A->5 , C->1
A->9 , C->0
A->4 , C->0.
495

Detailed Scratchpad

Input:
128+367
Target:
<scratch>
[1,2,8] has 3 digits.
[3,6,7] has 3 digits.
[1,2,8] + [3,6,7] , C=0, 8+7+0=15, A->5, C->1
[1,2] + [3, 6] , A= [5], 2+6+1=9 , A->9, C->0
[1] + [3] , A= [9,5] , C=0 , 1+3+0=4 , A->4 , C->0
[] + [] , A= [4,9,5] , C=0 , END
</scratch>
4 9 5



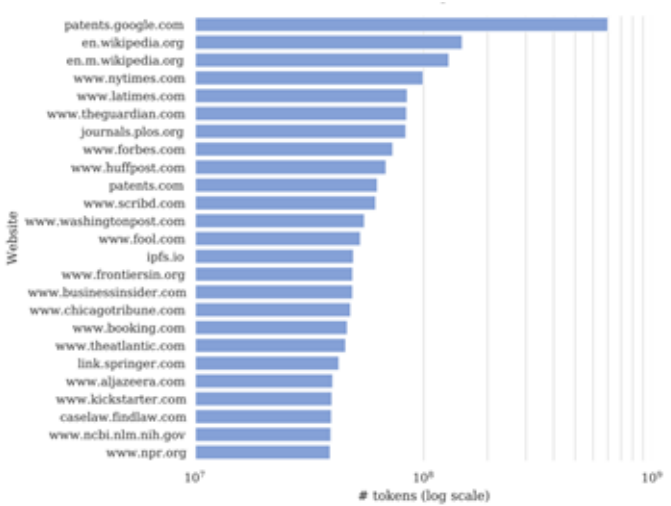
Data Filtering (1)

- As we saw, have to process data first
  - Filter out some points (toxicity, mismatch, etc)
  - Generally, we want “better” datasets
    - More diversity,
    - Less repeats.
- New benchmarks target this setting,
  - Fix the training procedure
  - Vary the data

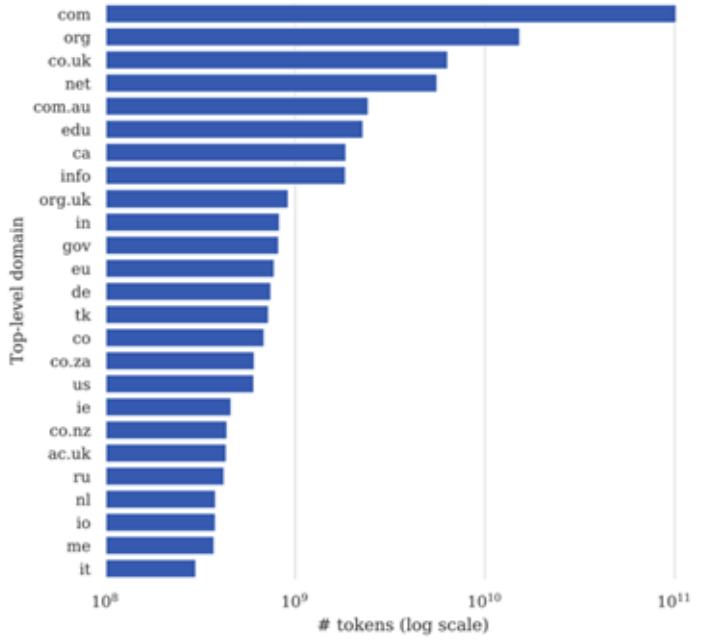
DataComp-LM: In search of the next generation of training sets for language models

Data Filtering (2)

- **Colossal Clean Crawled Corpus** (C4)
  - Removes bad words
  - Removes code
  - Language detection
  - ~800 GB (150 billion tokens)
  - Used to train T5 (Raffel et al ’23)
  - Analyzed by Dodge et al ‘21



Dodge et al ‘21



Redundancy Remove (1)

- **Document level redundancy**
  - Template string filter: Identifies and removes text containing excessive template content.

- **Sentence level redundancy**
  - N-gram repetition filter: Identifies repeated phrases of varying lengths and removes documents with excessive repetition, which may indicate poor content quality or artificial generation.
  - Semantic deduplication is particularly important for identifying paraphrased content, translated versions of the same material, and conceptually identical information

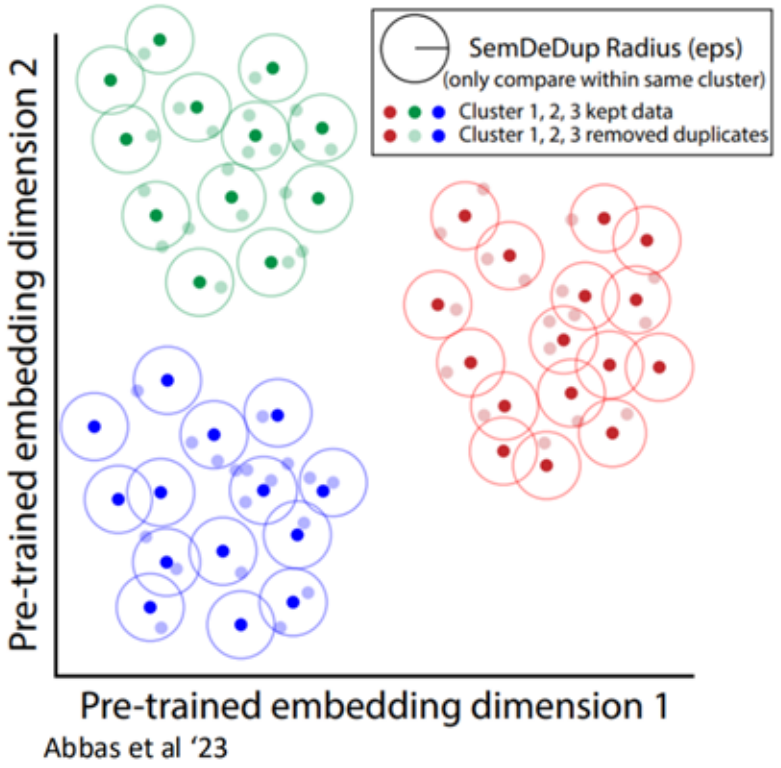
Redundancy Remove (2)

- "Deduplicating Training Data Makes Language Models Better ": Lee et al '22
  - Various ways to deduplicate data
  - Exact string matching
  - Approximate (hash-based, equivalent to embedding-based)

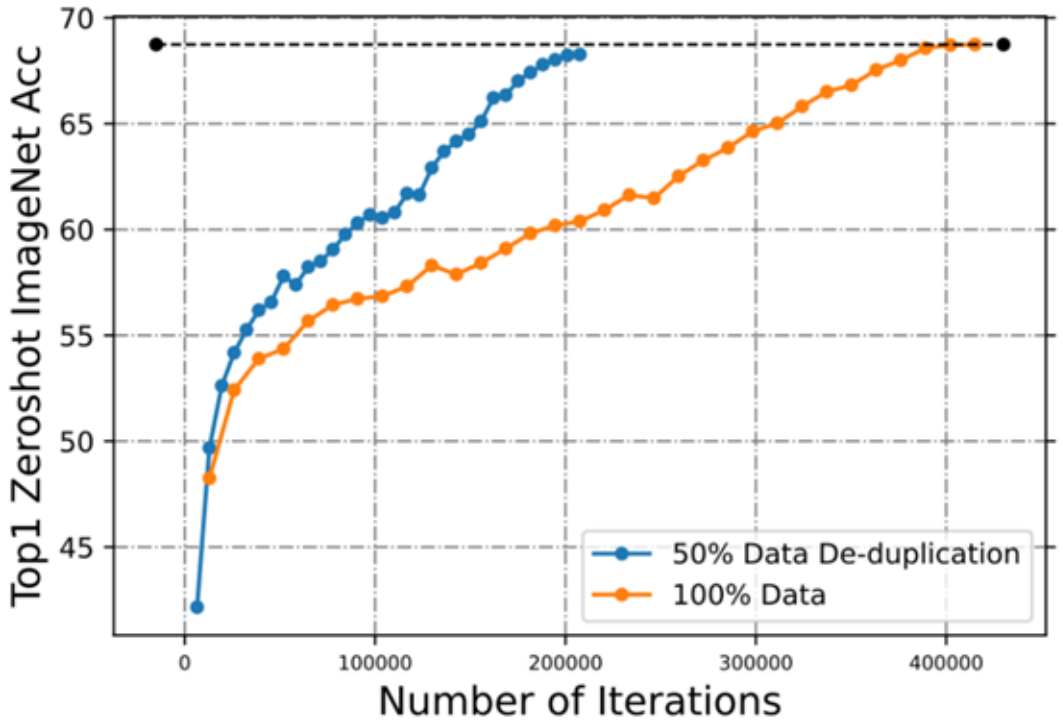
- One sentence shows up in **C4 60,000 times!**
  - "by combining fantastic ideas, interesting arrangements, and follow the current trends in the field of that make you more inspired and give artistic touches. We'd be honored if you can apply some or all of these design in your wedding. believe me, brilliant ideas would be perfect if it can be applied in real and make the people around you amazed!"

Redundancy Remove (3)

- How to define “duplicated” for data?
  - Idea: SemDeDup uses embeddings to identify near duplicates



Abbas et al ‘23



Model-based Data Selection (1)

- “High quality” data usually takes up a small percent
- How can we do data selection for low quality data

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

Model-based Data Selection (2)

- Use Wiki and Books data as the positive samples, and train a classifier (e.g. Bert)
- Use this classifier to select data from CC

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

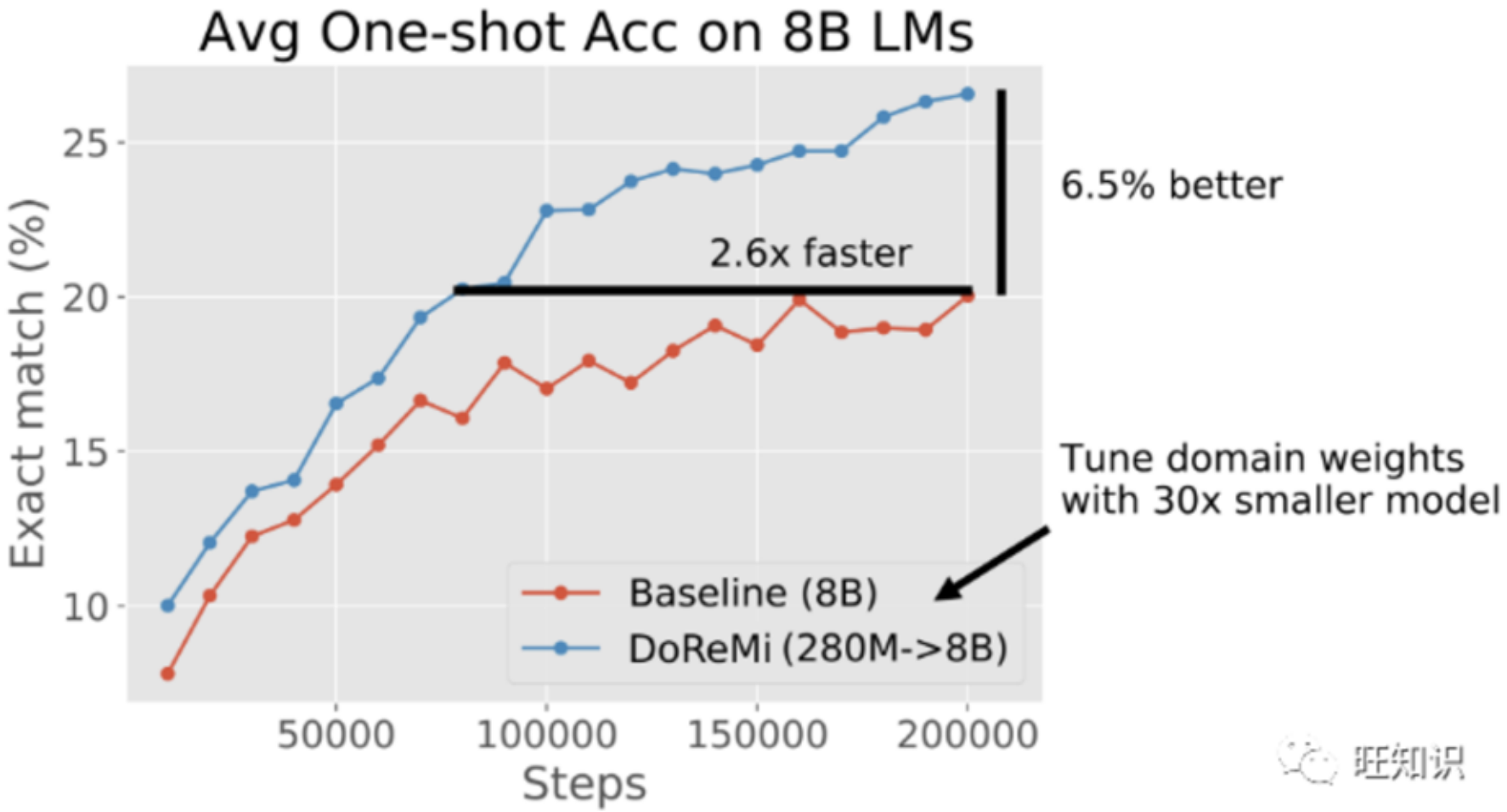
Model-based Data Selection (3)

- Llama 3 pretraining data
- Use Llama 2 as the classifier to select the data

- **Data Composition:** The 15T+ tokens consist of a new mix of publicly available online data, which includes four times more code than Llama 2.
- **Multilingual Data:** Over 5% of the pre-training dataset comprises high-quality non-English data covering over 30 languages, aimed at improving multilingual performance.
- **Quality Control:** The data was filtered using a combination of heuristic filters, NSFW filters, semantic deduplication, and text classifiers to ensure high quality.
- **Data Freshness:** The pre-training data has a cutoff of March 2023 for the 8B model and December 2023 for the 70B model.

How to Mix your data (1)

- Different mixing ratios may result in different learning speeds.



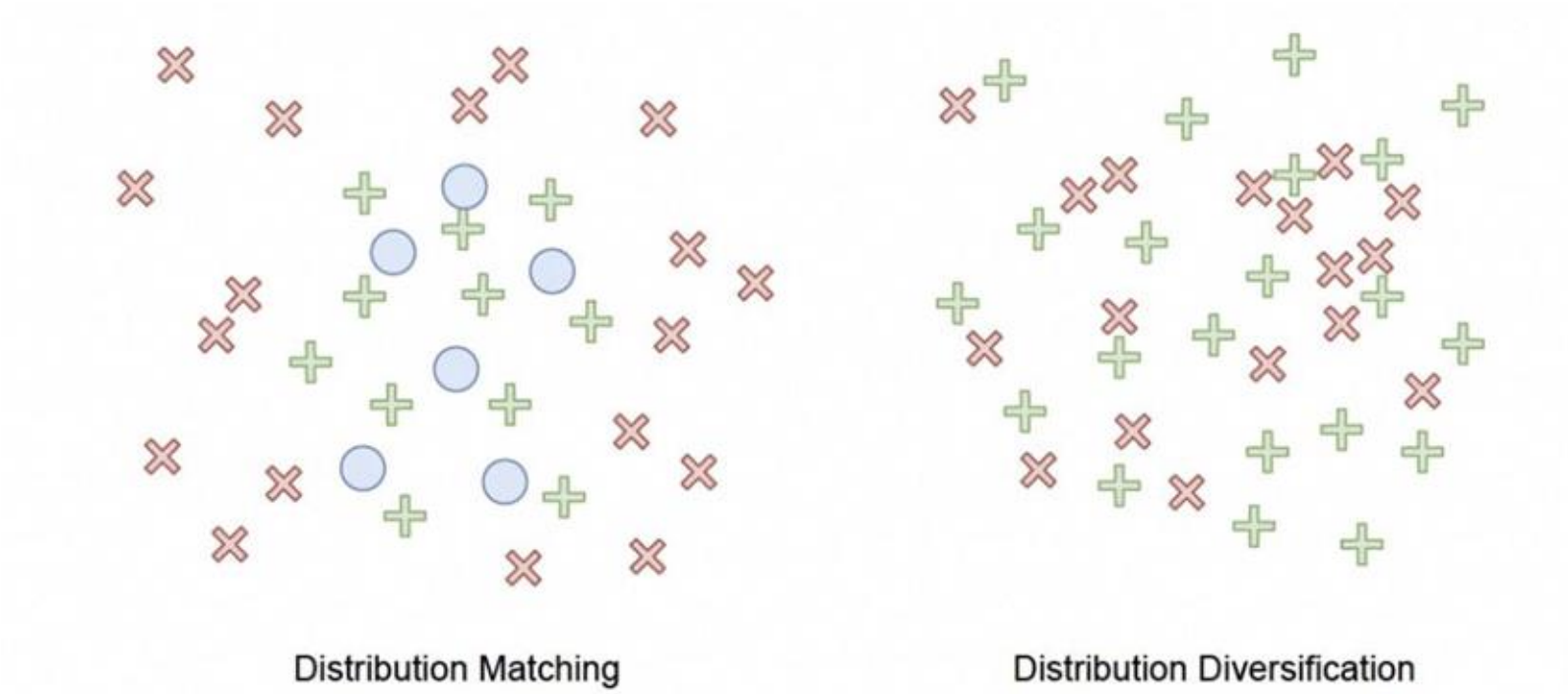
How to Mix your data (2)

- The code data is similar to that on GitHub, so LLaMA's coding performance isn't very high. In contrast, starcoder, which is primarily trained on code, performs well on coding tasks.
- The data is similar to that on papers like Arxiv, so LLaMA's scientific inference performance appears to be relatively low. In contrast, Galactica, which is primarily trained on papers, performs well in scientific inference.

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

How to Mix your data (3)

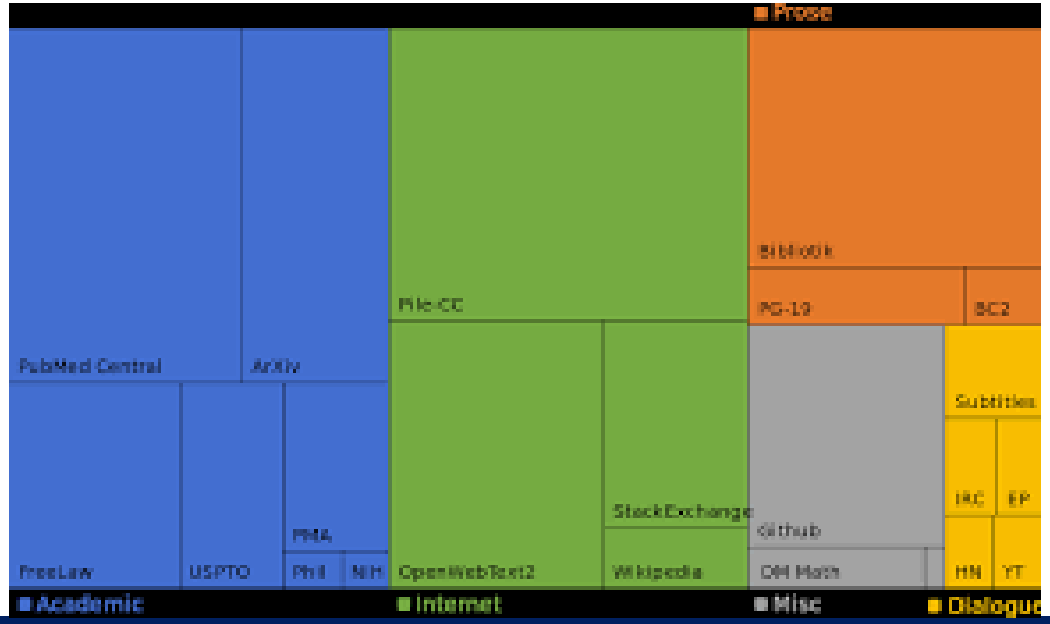
- Distributed learning methods prioritize sample heterogeneity and remove redundant data. The goal is to improve training efficiency and reduce bias.
- Diversity is required for both sample and category levels.



How to Mix your data (4)

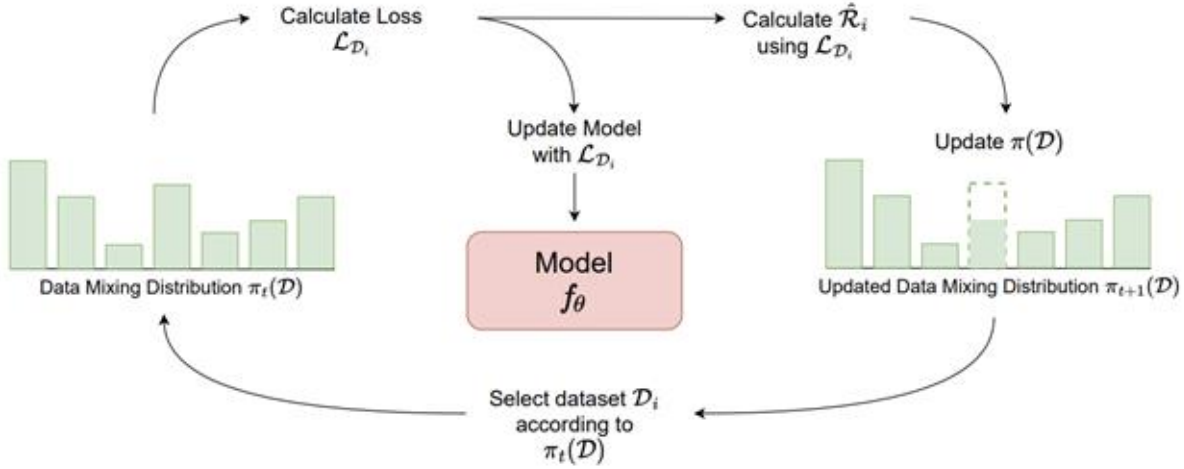
- Relying on intuition (or experiences): Improving the weights of domains with higher quality (potentially edited).

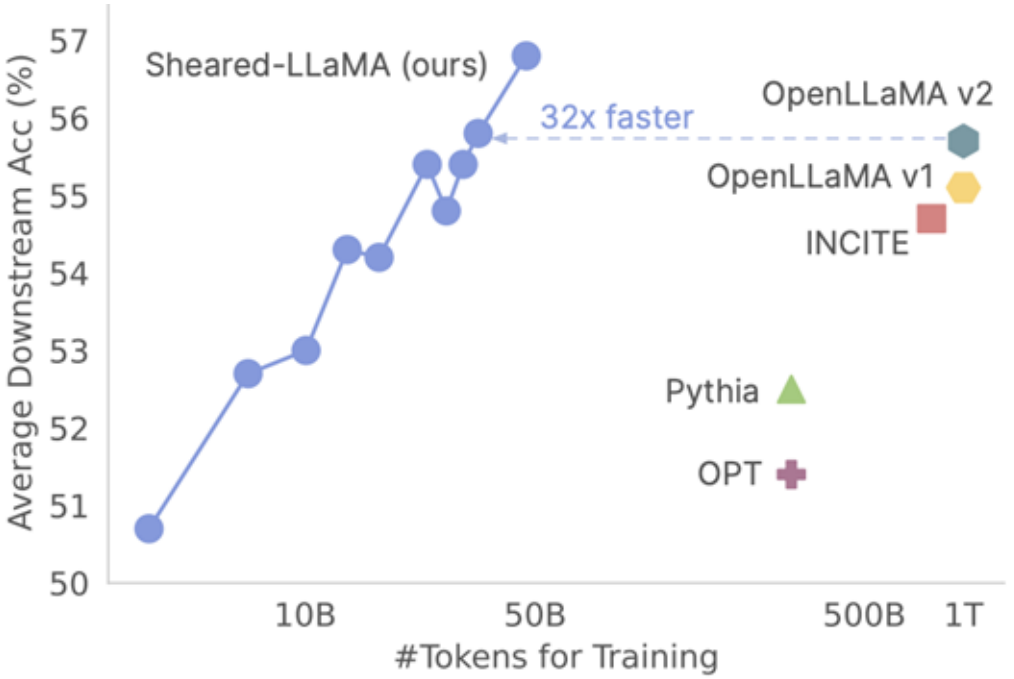
- Relying on downstream tasks: Adjusting domain weights based on performance on downstream tasks, training many models in zero-order optimization algorithms or heuristic search strategies.



How to Mix your data (5)

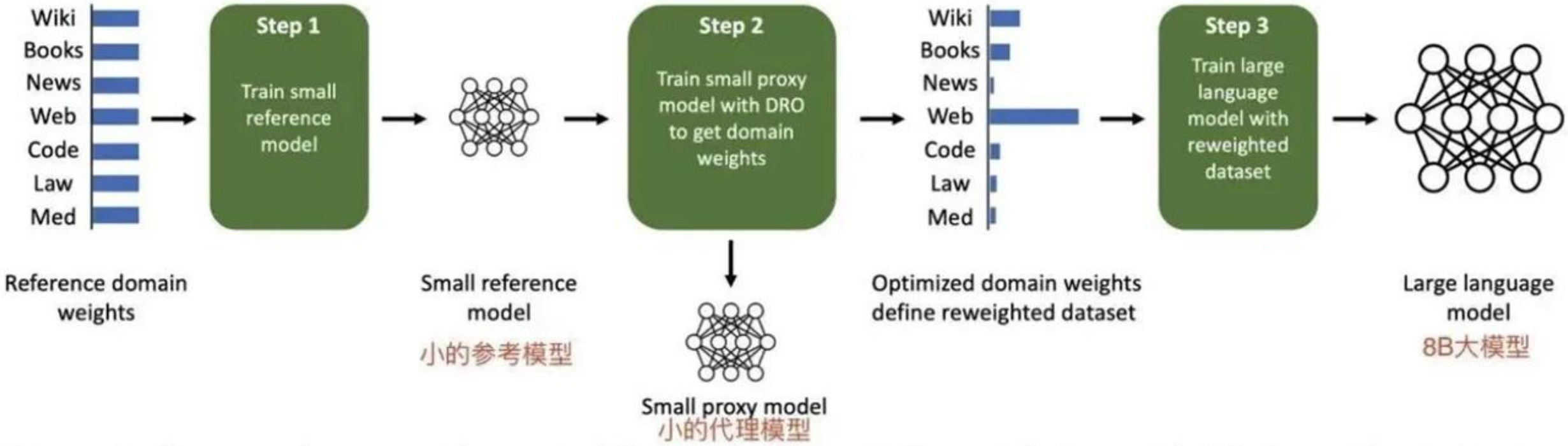
- **Online method**: Online Data Mixing (ODM) that combines elements from both data selection and data mixing, optimizing the data mixing proportions during training.





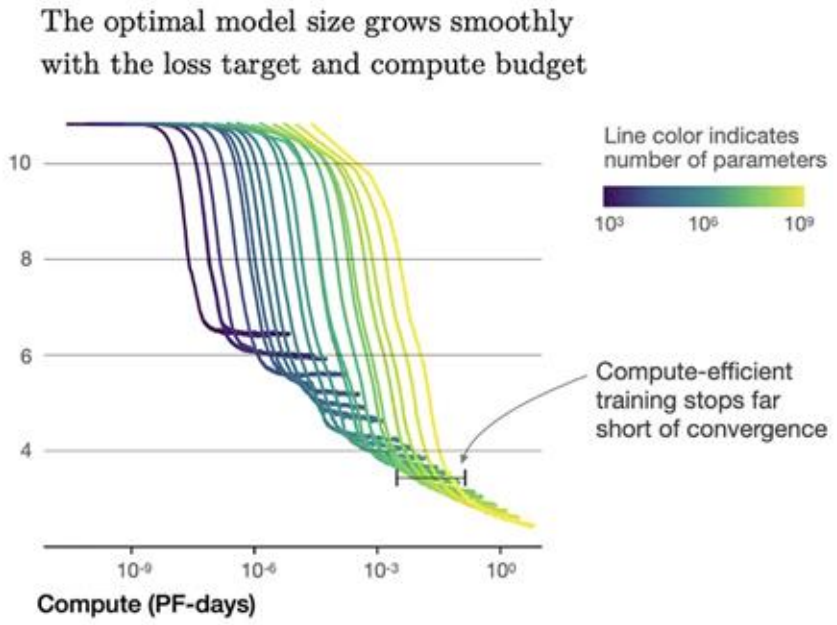
How to Mix your data (6)

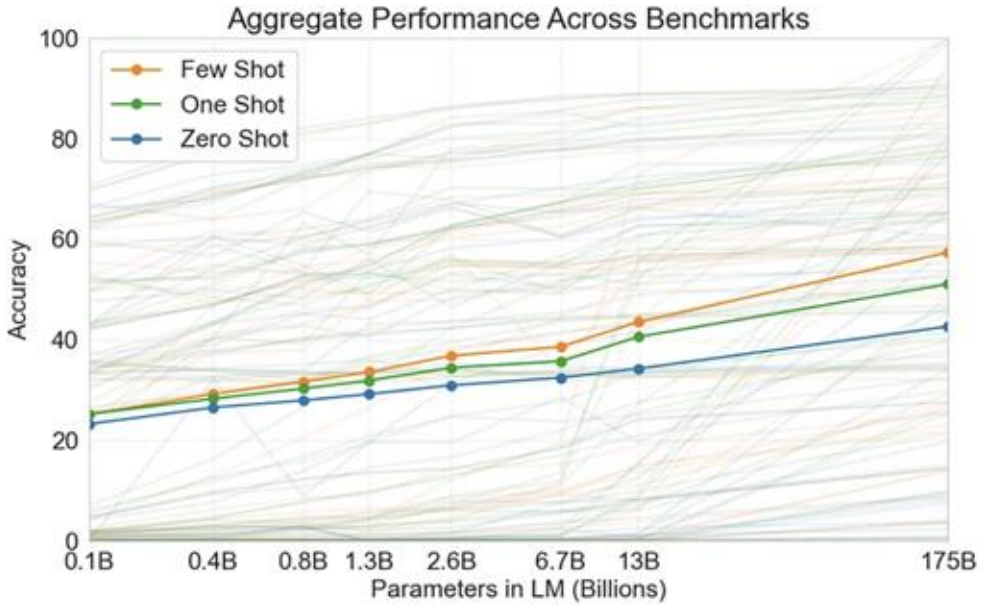
- **Offline method**: trains a small proxy model over domains to produce domain weights (mixture proportions) without knowledge of downstream tasks. Then resample a dataset with these domain weights and train a larger, full-sized model.



Model Scaling (1)

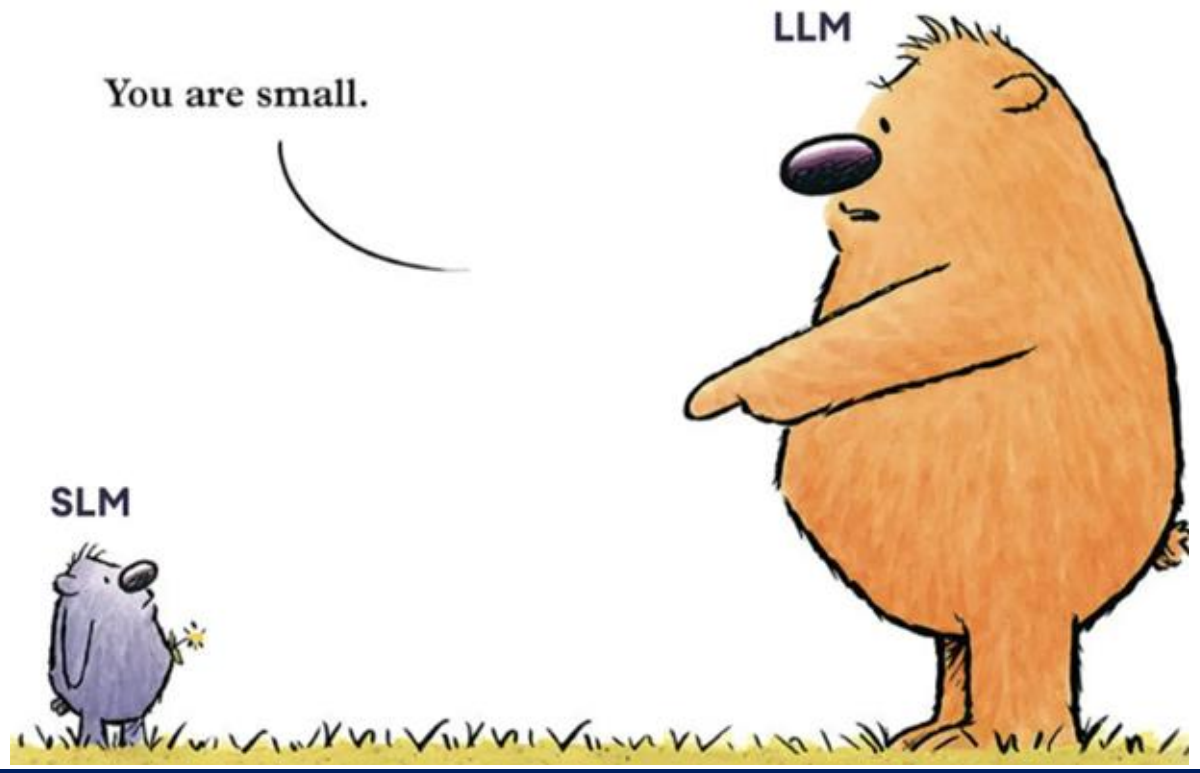
- Can we use the results from small models to guide the design of large models?
- Yes, we have **Scaling Laws.** OpenAI reports that during the development of GPT-4, they predicted the performance of HumanEval before the experiment.





Model Scaling (2)

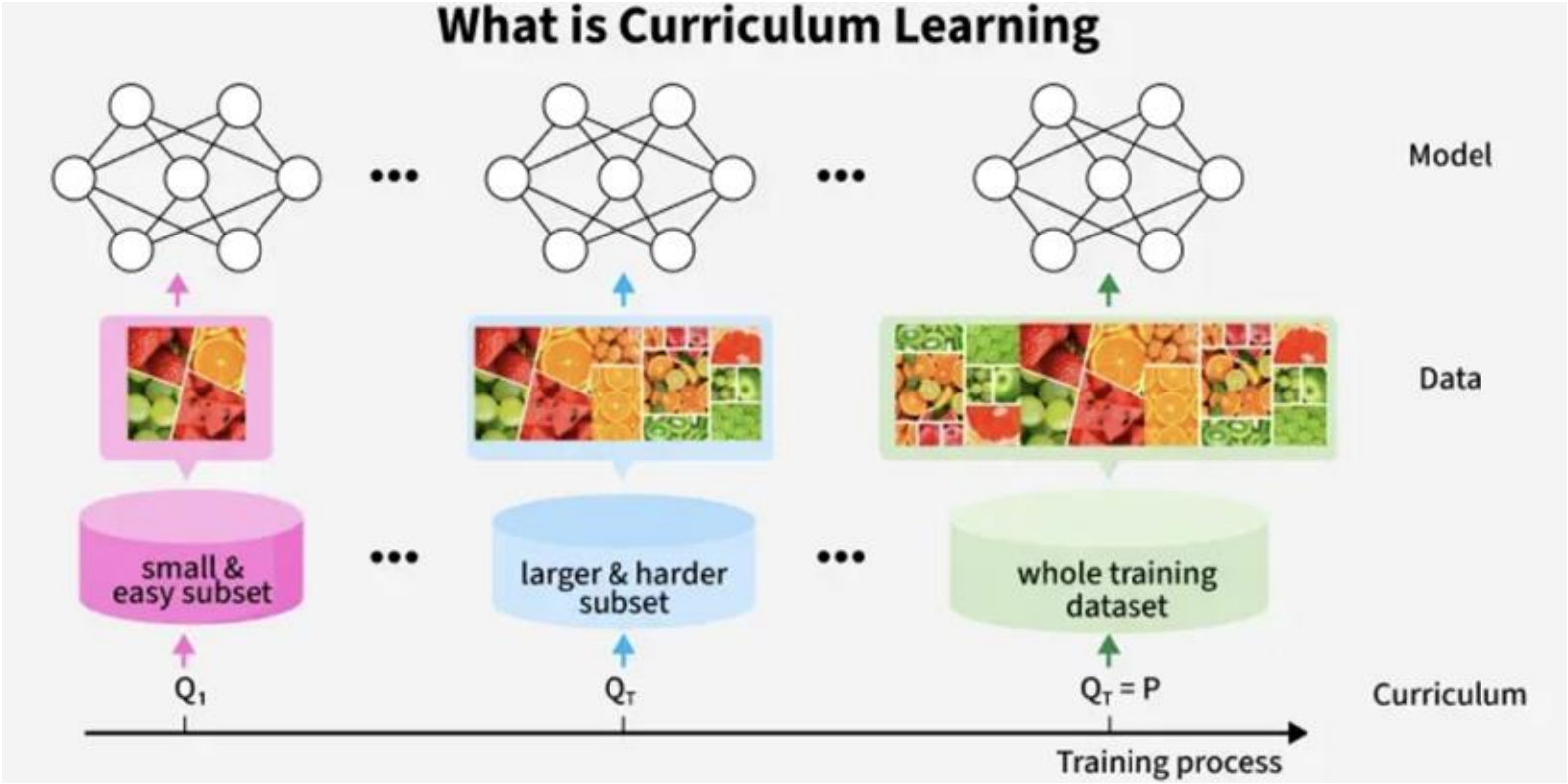
- Can we use the results from small models to guide the design of large models?
- No, some observations can not be transferred from small models to large models. (e.g., 7B vs. 70B)



Data Curriculum (1)

- We aim to teach the model many skills, such as math and coding.

- Different training data sources exist to aid each skill (e.g., web pages for MMLU-type problems, GitHub resources for coding problems).

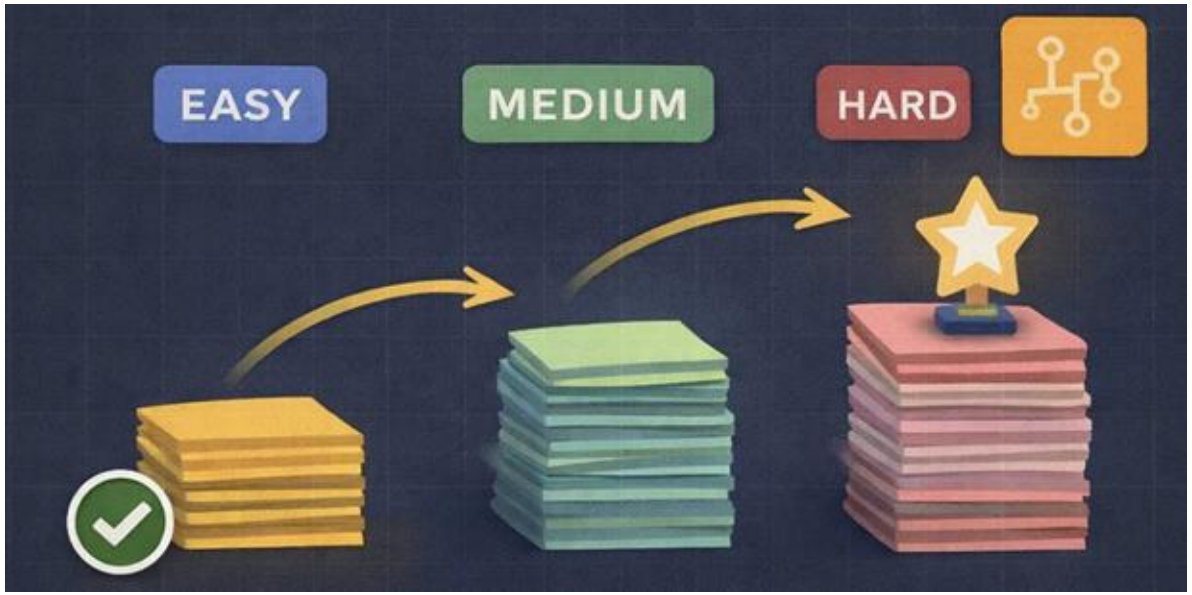


Data Curriculum (2)

- Method 1 (Code Only): Directly feed 10B code data

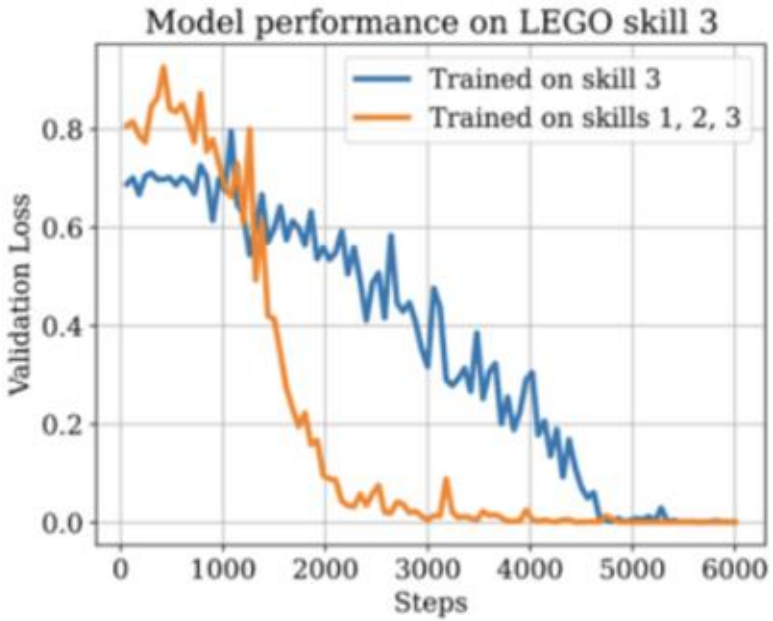
- Method 2 (Uniform Mixing): Uniformly mix 5B text and 5B code data, then input them into the model simultaneously.

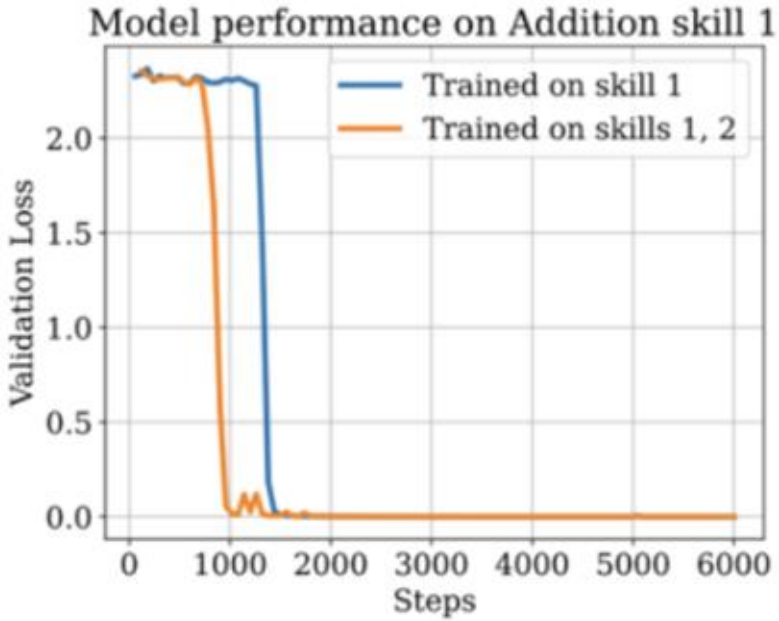
- Method 3 (Data Curriculum): Input 5B text first, then input 5B code.

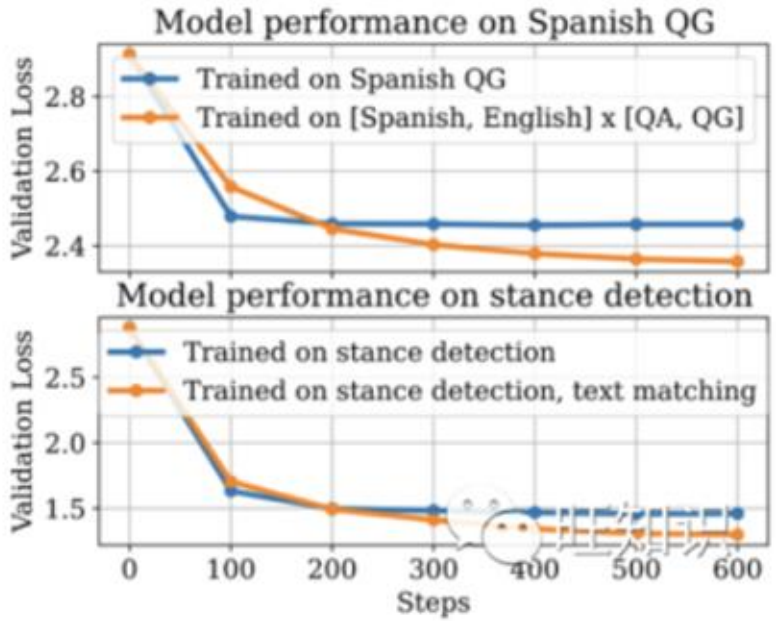


Data Curriculum (3)

- If the skills learned by the model from text data are not helpful for code data, then we can use 1, as in the cases of StarCoder and AlphaCode.
- If the skills learned by the model from text data can be transferred to code data, then we might want to do 2, uniform mixing.
- If learning code skills requires the model to already have text skills, and the text must come first—then we need to do 3, data-based learning. Codex and CodeLLaMA are examples of this.







Places to Get Data: RedPajama

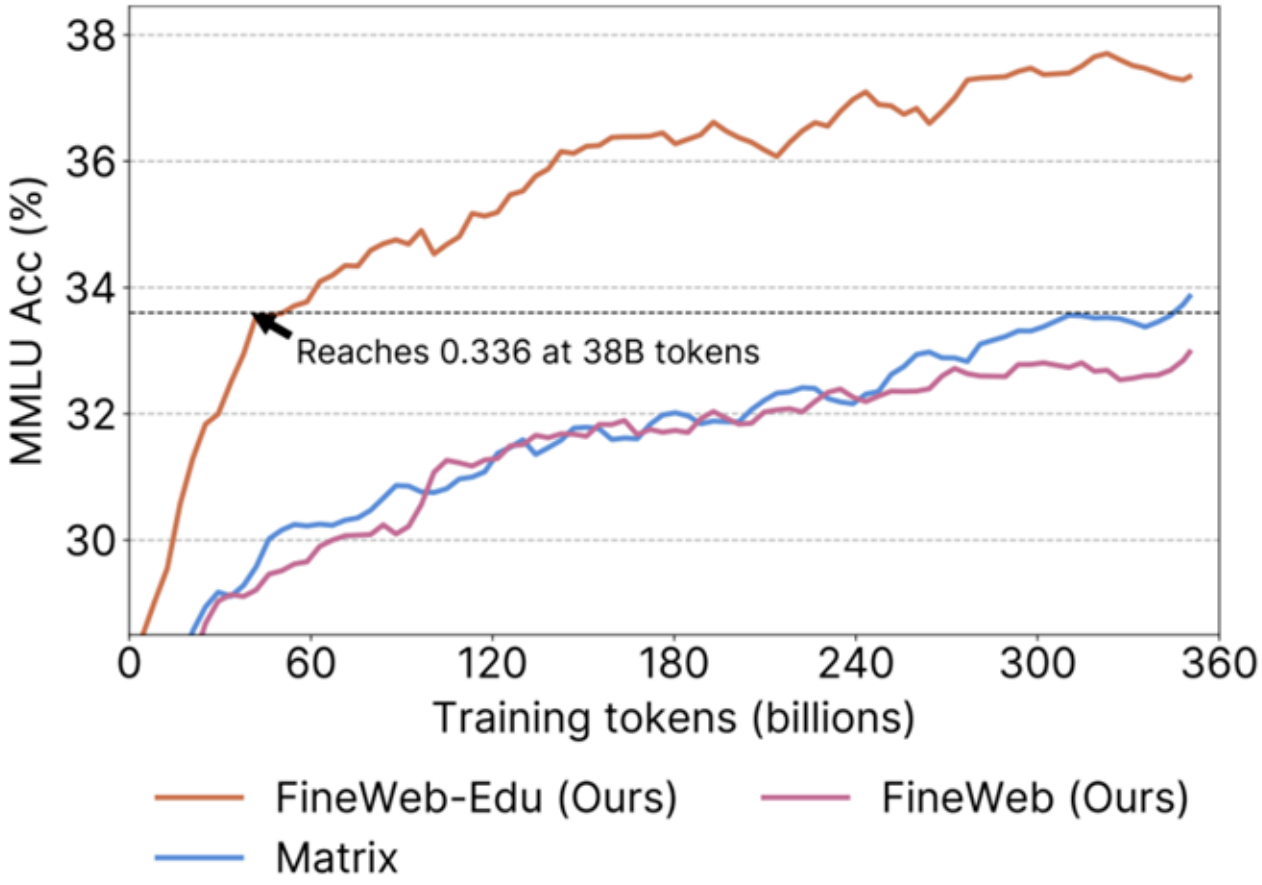
- **RedPajama v2**
  - Open dataset with 30 trillion tokens
  - Oct ’23 / Together AI
  - Pre-computed quality annotations
    - “ML classifiers on data quality, minhash results that can be used for fuzzy deduplication, or heuristics such as “the fraction of words that contain no alphabetical character”. ”

	# Documents	Estimated Token count (deduped)
en	14.5B	20.5T
de	1.9B	3.0T
fr	1.6B	2.7T
es	1.8B	2.8T
it	0.9B	1.5T
Total	20.8B	30.4T

github.com/togethercomputer/RedPajama-Data

Places to Get Data: FineWeb

- **FineWeb**
  - Open dataset with 15 trillion tokens
  - June ’24 / Hugging Face
  - Additional filtered ”educational” data
  - Full data construction and experimental details available.



Scraping from the Web

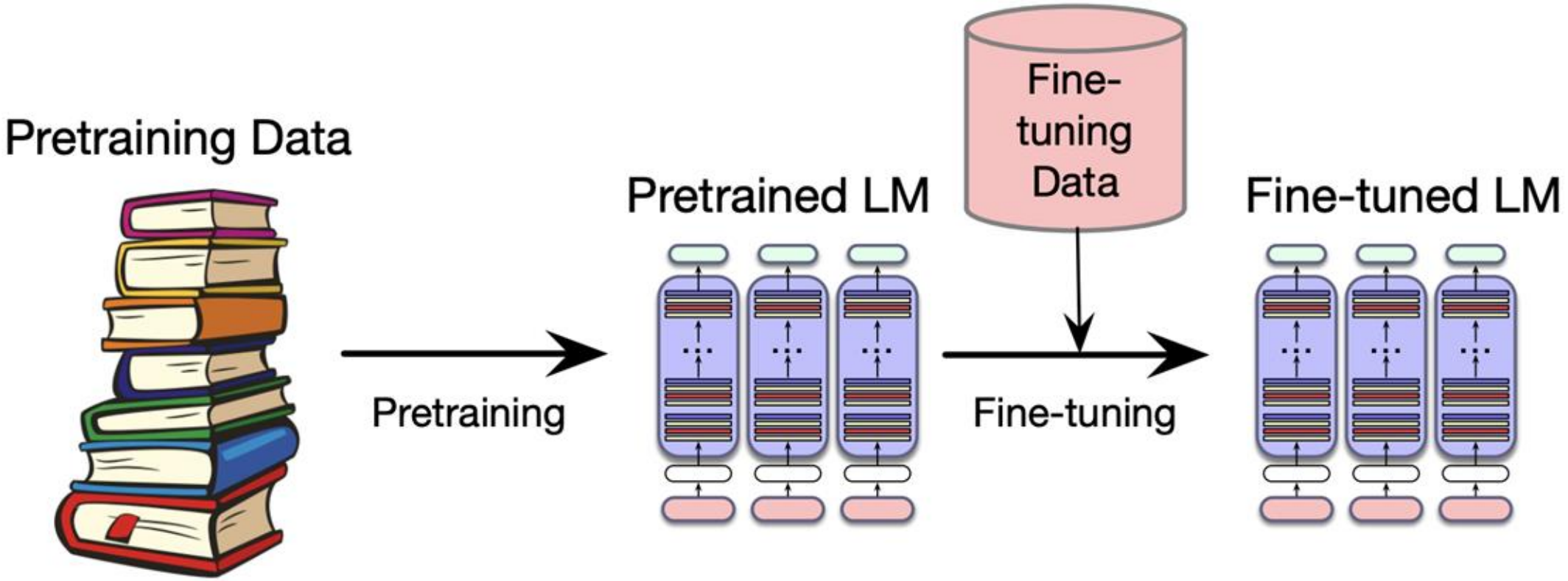
- **Copyright: much of the text in these datasets is copyrighted**
  - Not clear if fair use doctrine in US allows for this use
  - This remains an open legal question

- **Data consent**
  - Website owners can indicate they don't want their site crawled

- **Data Privacy**
  - Websites can contain private IP addresses and phone numbers

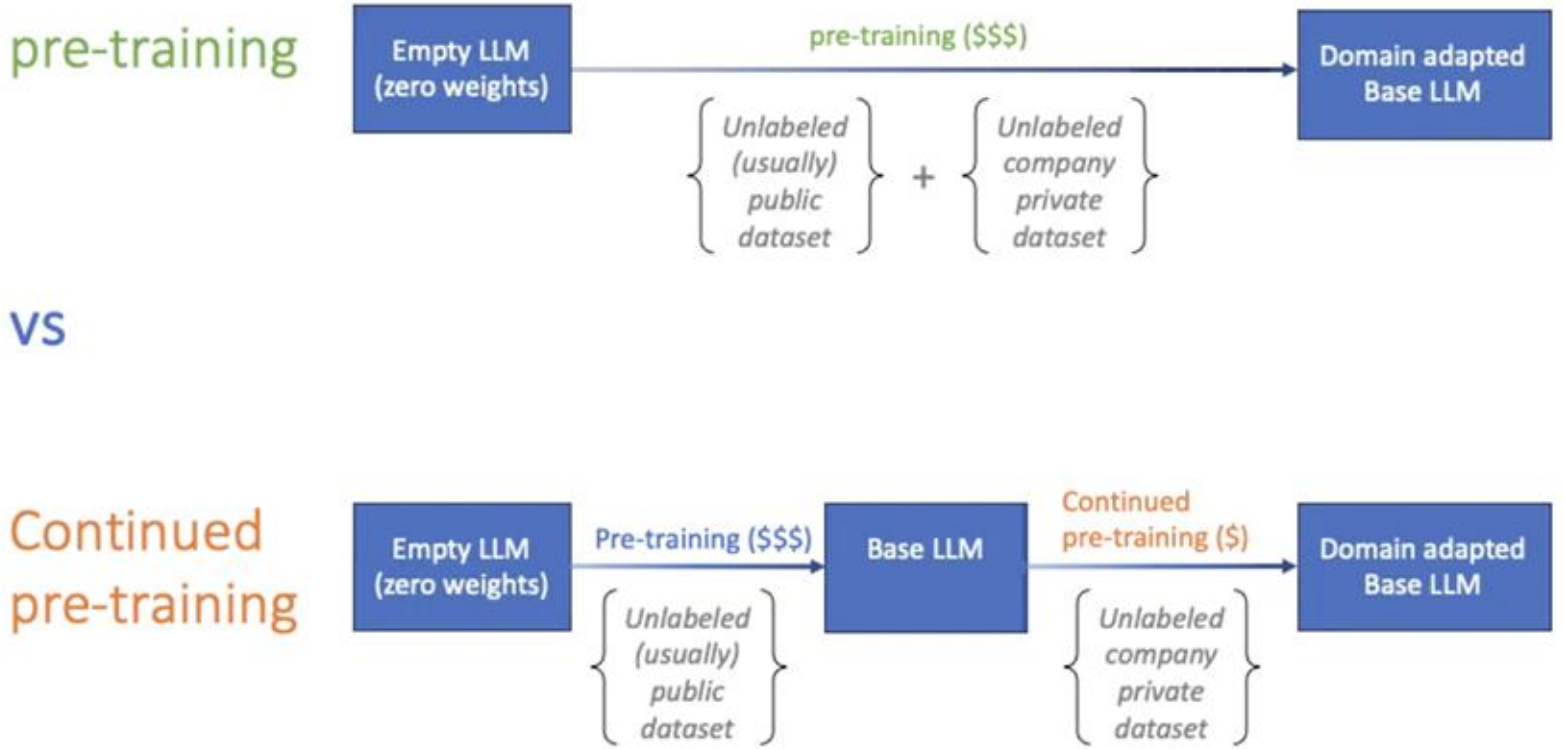
Fine-tuning (Post-training)

- What happens if we need our LLM to work well on a domain it didn't see in pretraining? Perhaps some specific medical or legal domain?



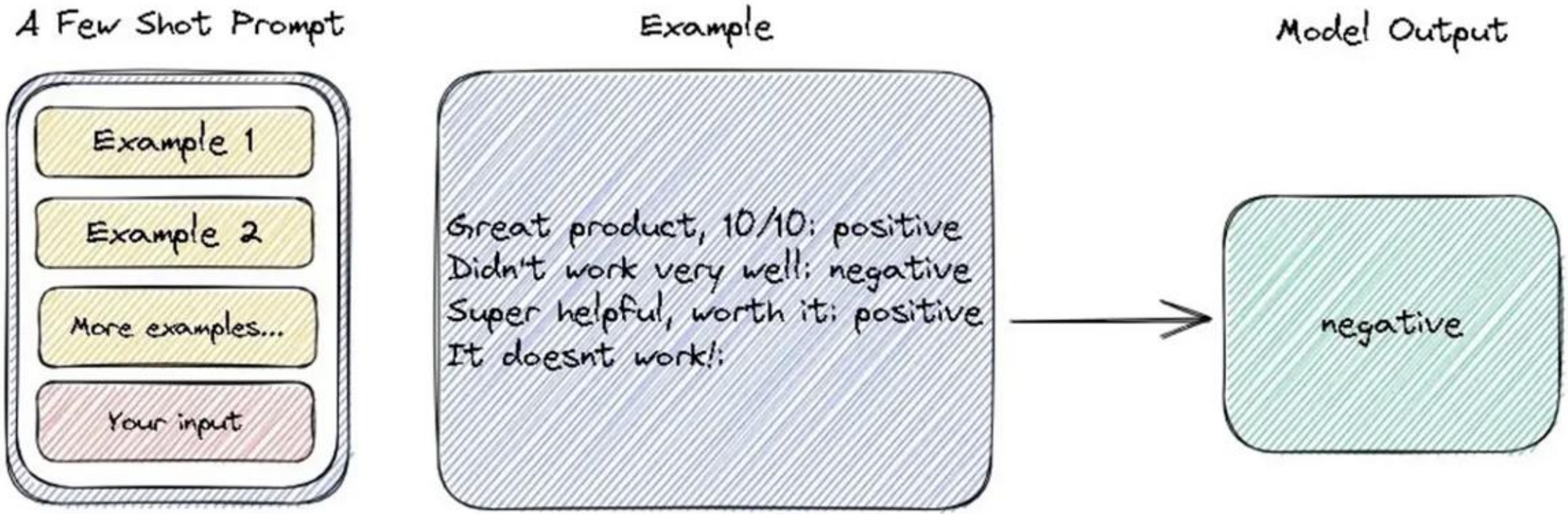
Fine-tuning (Post-training)

- Taking a pretrained model and further adapting some or all of its parameters to some new data
- Hence sometimes called continued pretraining



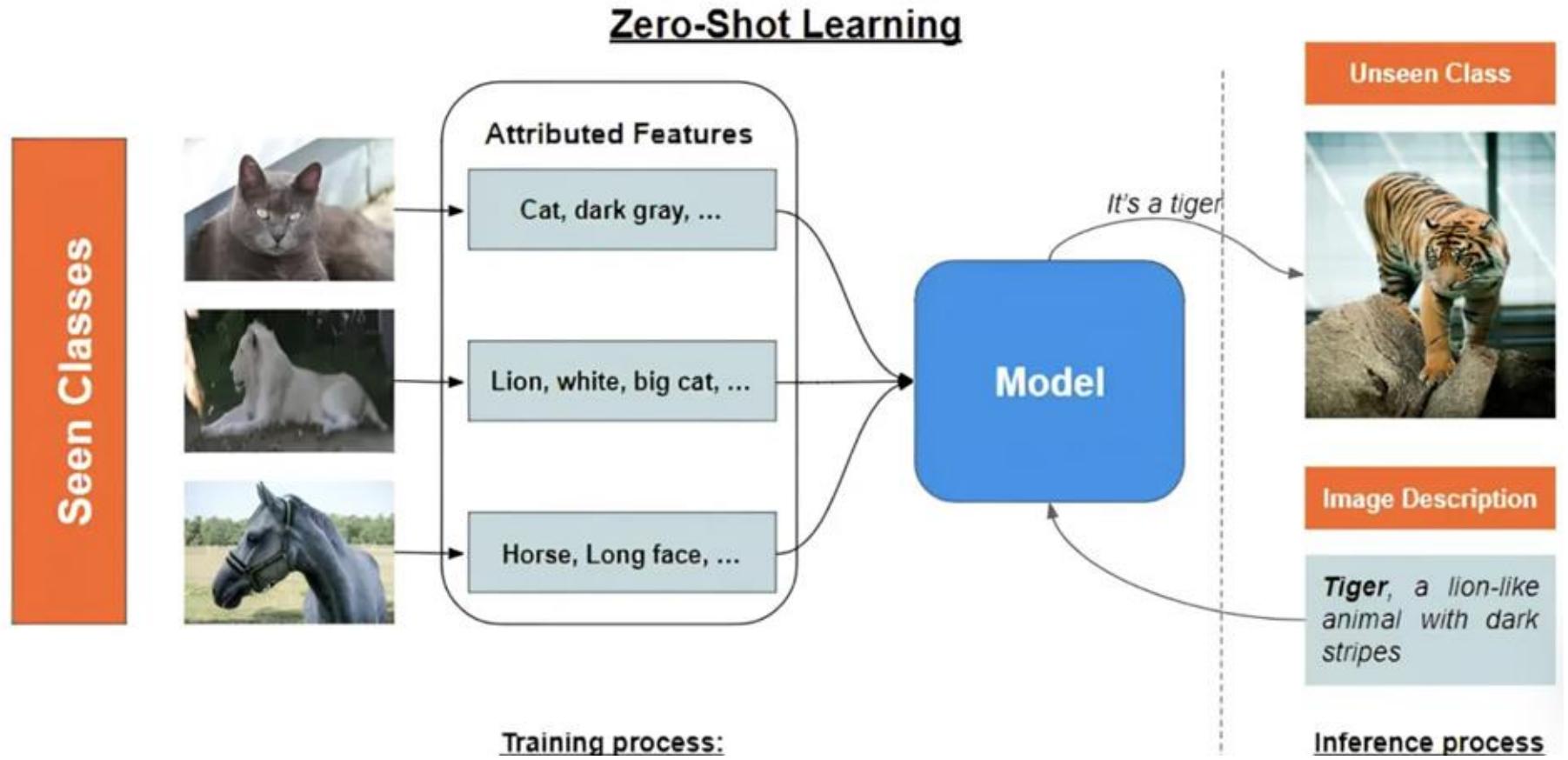
Few-shot Learning

- A machine learning approach that enables a model to generalize and make accurate predictions after being trained with only a small number of labeled examples per class



Zero-shot Learning

• Zero-shot learning (ZSL) is an AI technique where models classify or understand new data/tasks they weren't explicitly trained on.



Where do we use FT?

- Vertical Enhancement: Enhances the general-purpose expertise of large models, such as in healthcare, law, and finance.

- Task Specialization: Optimizes for specific scenarios and tasks, such as question answering systems, text classification, and named entity recognition.

- Language Localization: Adapts models primarily trained on English or Chinese to other languages.

- Tone and Style Adjustment: Adjusts the model's output style to better match specific expressions or tones.

- Security Enhancement: Reduces the likelihood of the model generating harmful or inappropriate content through fine-tuning.

How to collect FT data?

- **Web data + manual rewriting**: after the data is crawled, it usually needs to be manually reviewed and rewritten to ensure data quality and applicability.

- **Production data + manual rewriting**: data may come from closed-source LLM responses and may lack the characteristics of real human conversations.

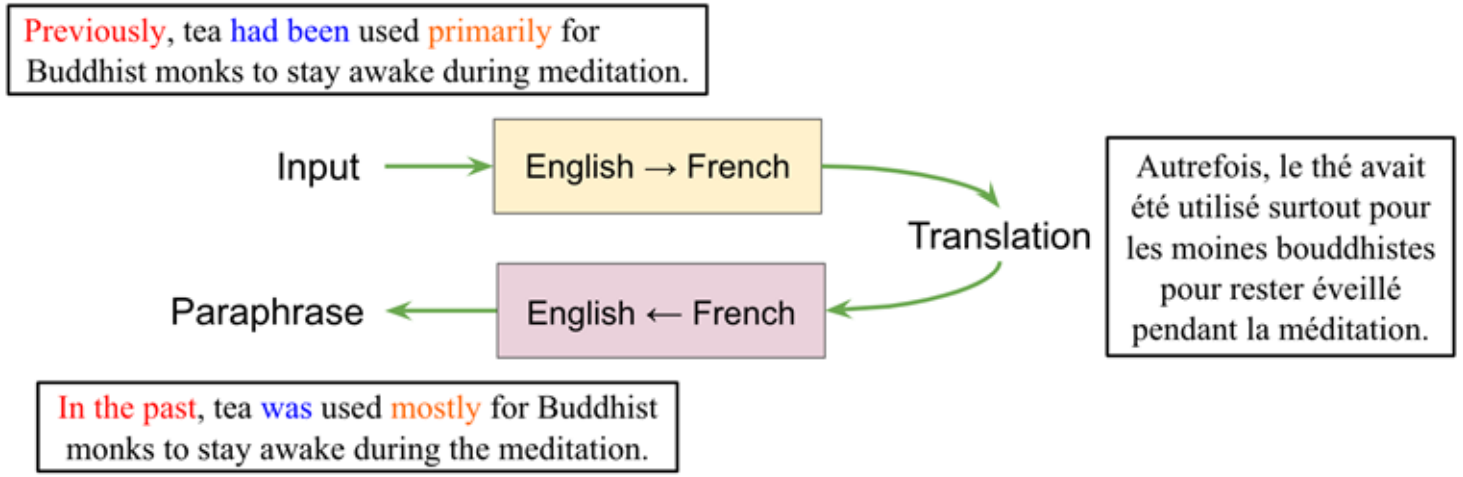
- **Social media data + manual rewriting**: when exporting to social media platforms, it is important to handle issues involving privacy information (PII), such as name, ID number, and phone number.

- **Public dataset**: By making good use of these public datasets, we can provide the model with rich role-playing training materials, thereby improving the model's dialogue ability and role performance.

FT Data Augmentation (1)

• Synonym Substitution: Using a thesaurus or pre-trained word embedding models, replace certain words in a sentence. This method can increase linguistic diversity while preserving the original semantics.

• Back Translation: Translate the original text into another language and then back. This method can generate different sentence structures expressing the same meaning.



FT Data Augmentation (2)

• Contextual Expansion: Add more contextual information to existing dialogues, making them richer and more natural.

• Sentence Restructuring: Break long sentences into shorter sentences, or combine multiple short sentences into longer sentences, increasing the diversity of sentence structures.

• Text Generation: Utilize existing large language models to generate new and similar text. This method can generate a large amount of new data, but it requires careful control of the quality and relevance of the generated content.

Data Augmentation with LLMs (1)

content expansion

Original Dialogue: User: I'm not feeling well today.

Assistant: I'm sorry to hear that. Could you be more specific about where it hurts?

Expanded Dialogue:

User: I'm not feeling well today.

Assistant: I'm sorry to hear that. Could you be more specific about where it hurts?

User: I feel dizzy and a little nauseous.

Assistant: It sounds like you might be experiencing dizziness and nausea. These symptoms could have several causes. Have you been staying up late or under a lot of stress lately? Or did you eat anything unsuitable? User: Yes, I've been under a lot of work stress lately, and I often stay up very late.

Assistant: I understand. Work stress and lack of sleep can indeed cause these symptoms. I suggest you try adjusting your sleep schedule and ensuring you get enough sleep. Also, try to relax your mind and body; you can try some simple meditation or deep breathing exercises. If the symptoms persist or worsen, it's best to see a doctor. Do you think these suggestions are feasible?

Data Augmentation with LLMs (2)

scenario change

Original dialogue (in a coffee shop):

User: The coffee here is really good.

Assistant: Yes, they use select Arabica beans, and the taste is truly excellent.

We can have LLM transform this dialogue into different scenarios, such as a restaurant, a bookstore, etc.

Transformed dialogue (in a restaurant):

User: The steak here is really good.

Assistant: That's right, they use select Angus beef, and the taste is truly excellent.

Data Augmentation with LLMs (3)

style transfer

For example: Original dialogue (formal style): User: Could you recommend a book on artificial intelligence?

Assistant: Sure. I suggest you consider reading *Artificial Intelligence: A Modern Approach* by Stuart Russell and Peter Norvig. It's a classic textbook that comprehensively introduces the field of AI.

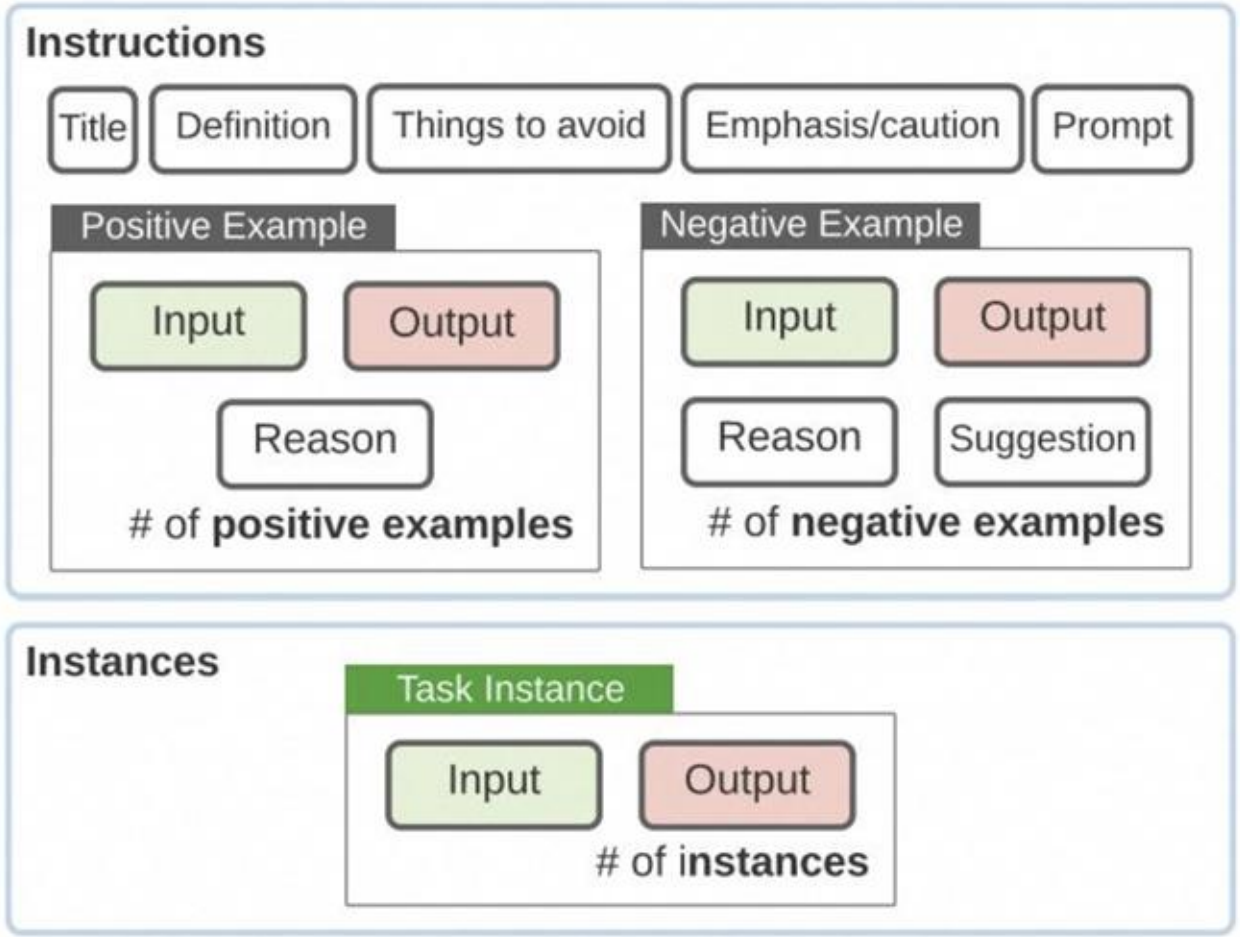
Converted dialogue (relaxed style):

User: Hey, any good AI books to recommend?

Assistant: Wow, speaking of good AI books, one that immediately comes to mind is amazing! *Artificial Intelligence: A Modern Approach* by Russell and Norvig is a treasure trove. It covers AI knowledge so comprehensively; it will definitely broaden your horizons!

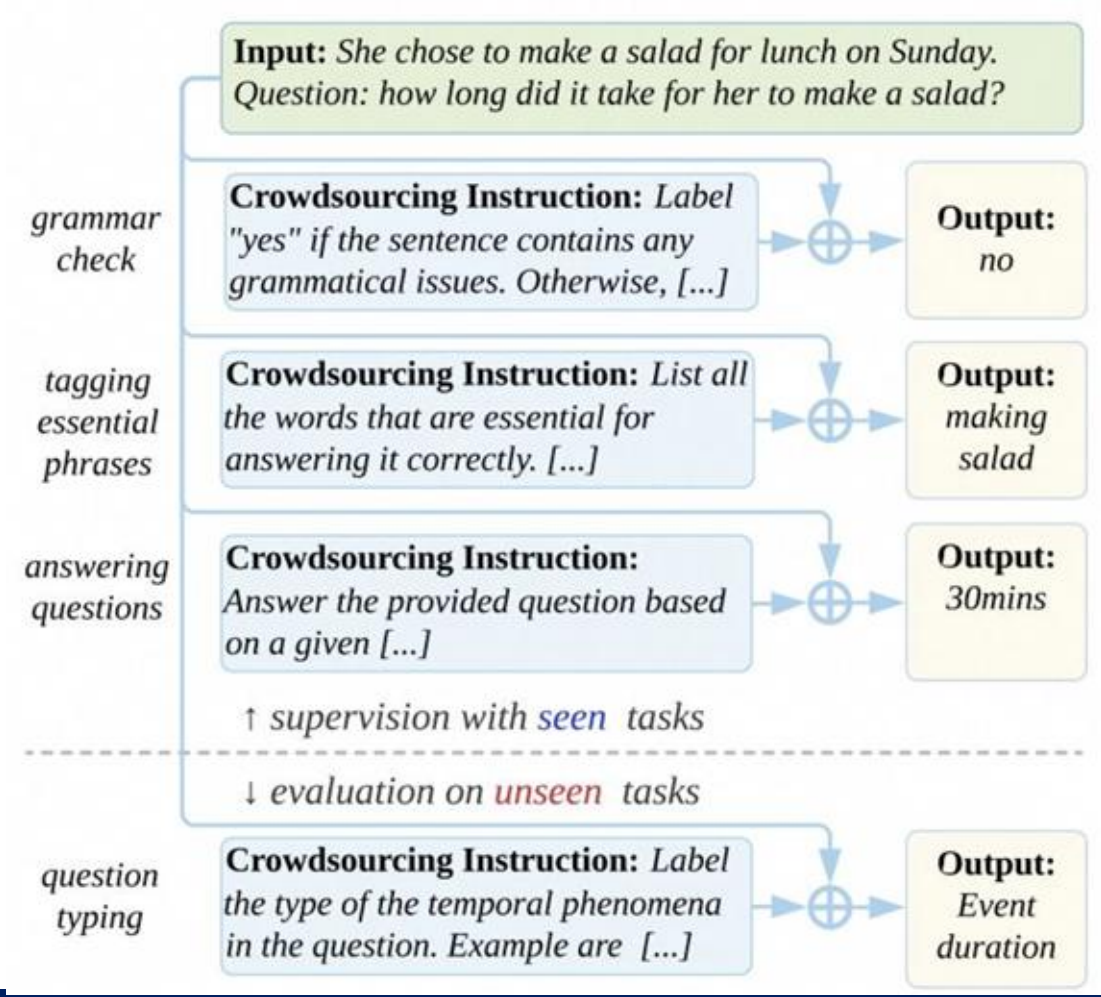
Instruction Data (1)

• The purpose of instruction fine-tuning is to enable the model to better understand and execute human-provided instructions, to significantly improve zero-shot capability for unseen tasks.



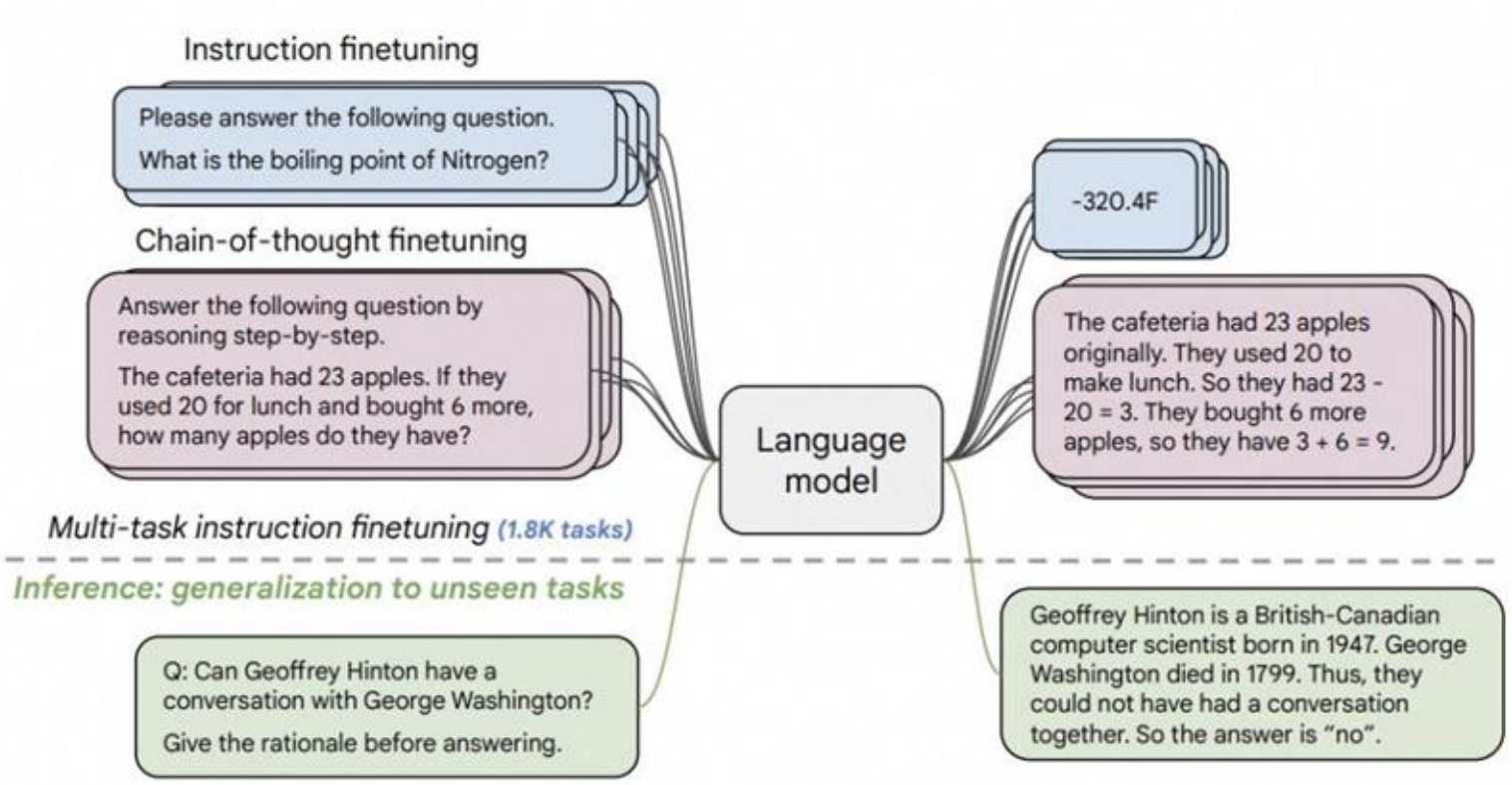
Instruction Data (2)

- The language model performs supervised fine-tuning on this data, taking the task instructions and input as inputs to the language model to generate the corresponding output.



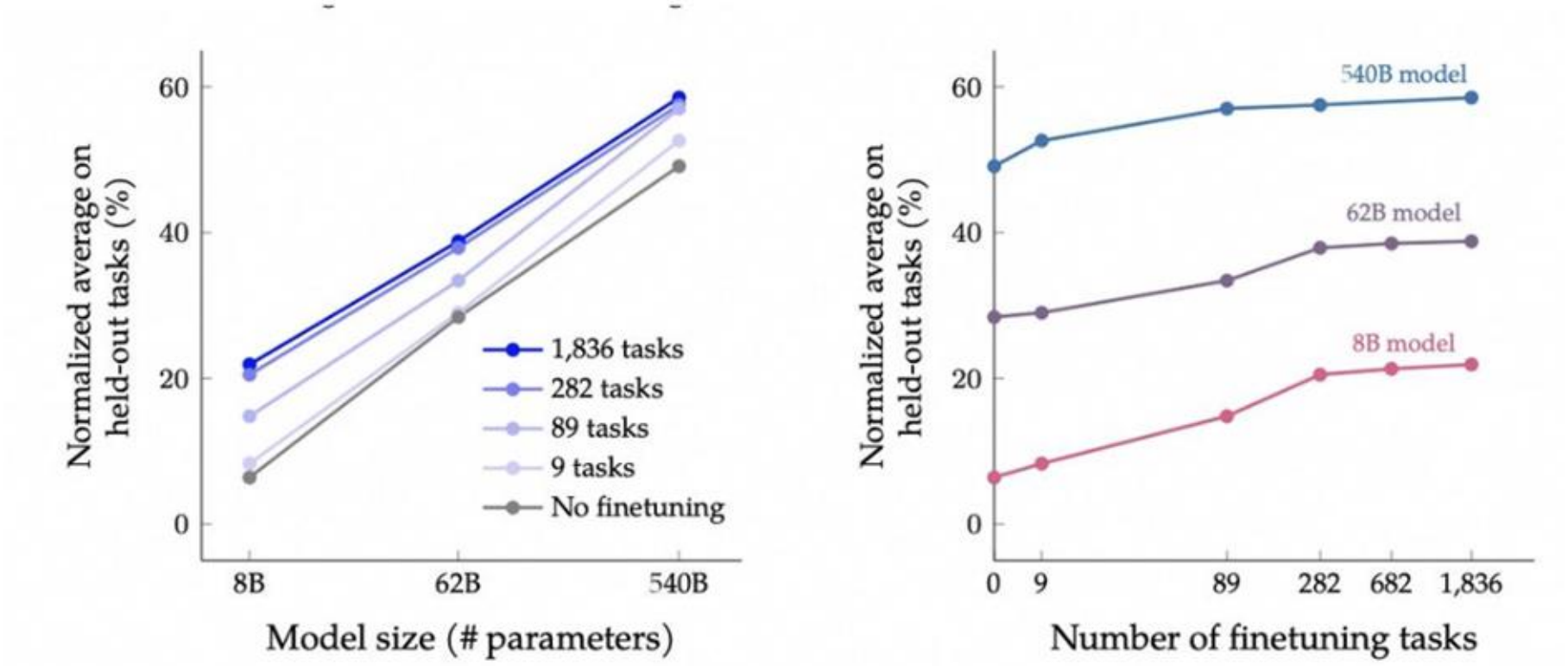
Instruction Data (3)

- The language model performs supervised fine-tuning on this data, taking the task instructions and input as inputs to the language model to generate the corresponding output.



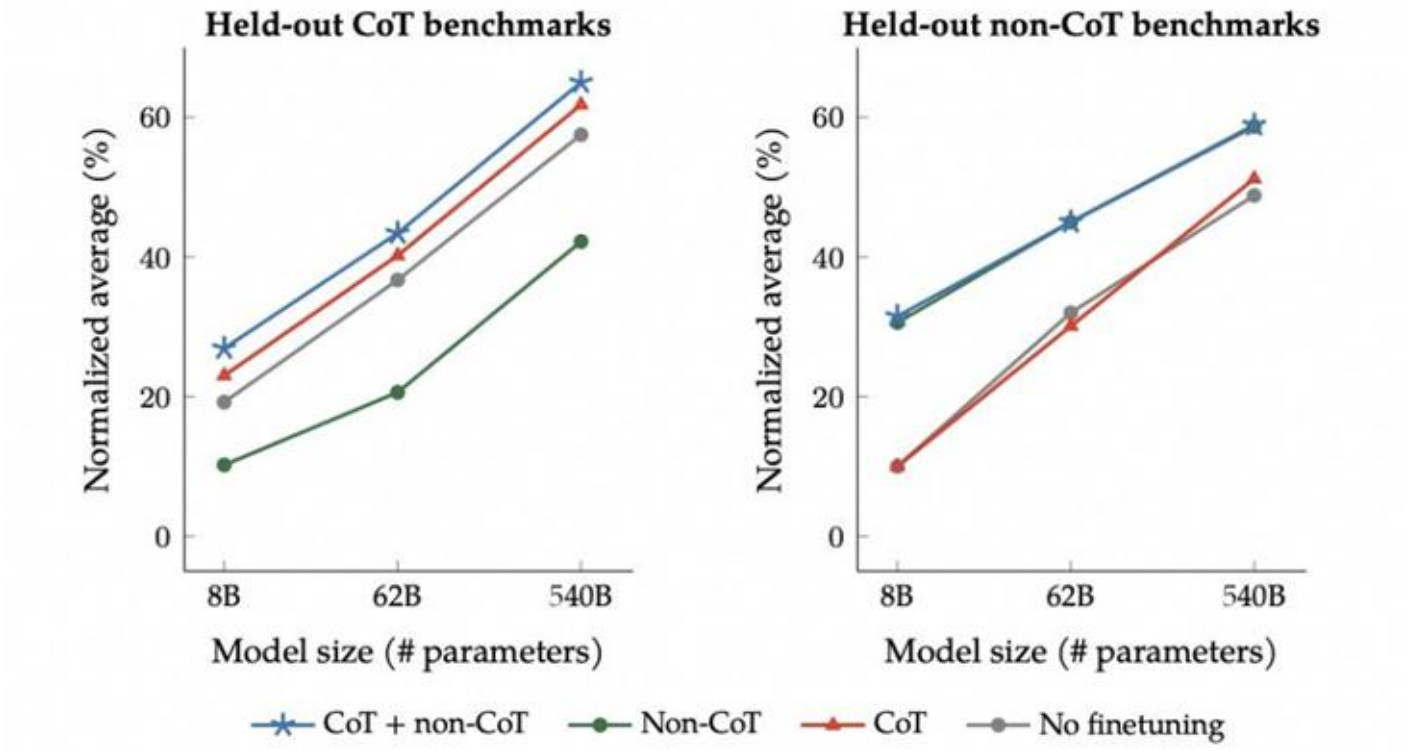
Instruction Data (4)

• As the model size and the number of instruction tuning tasks increase, the model performance gradually improves. When the number of tasks exceeds 282, the model performance does not change significantly.



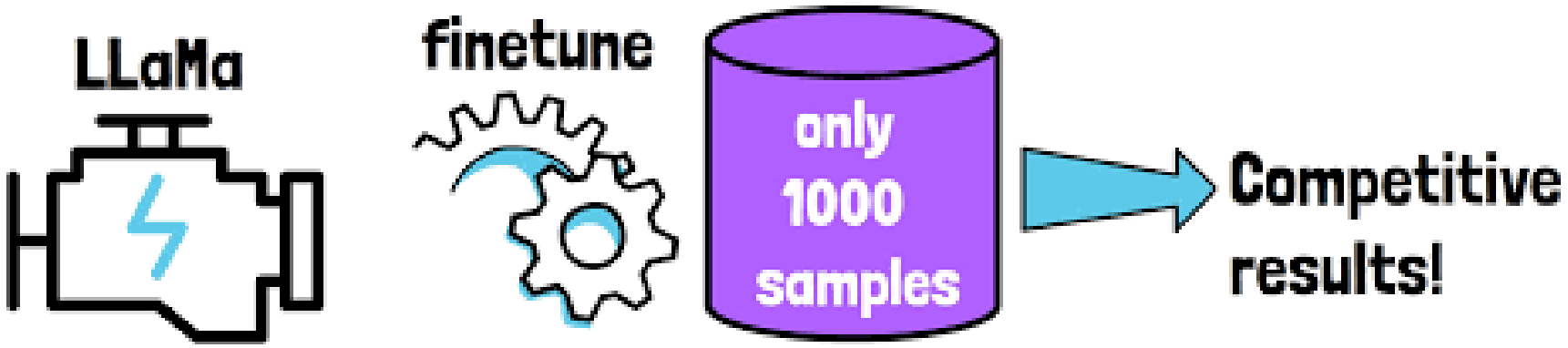
Instruction Data (5)

•Including COT data can improve model inference capabilities and further unlock zero-shot capabilities. Furthermore, instruction tuning without COT data will negatively impact performance on COT datasets, while instruction tuning with a mix of COT and non-COT data will result in models that perform well.



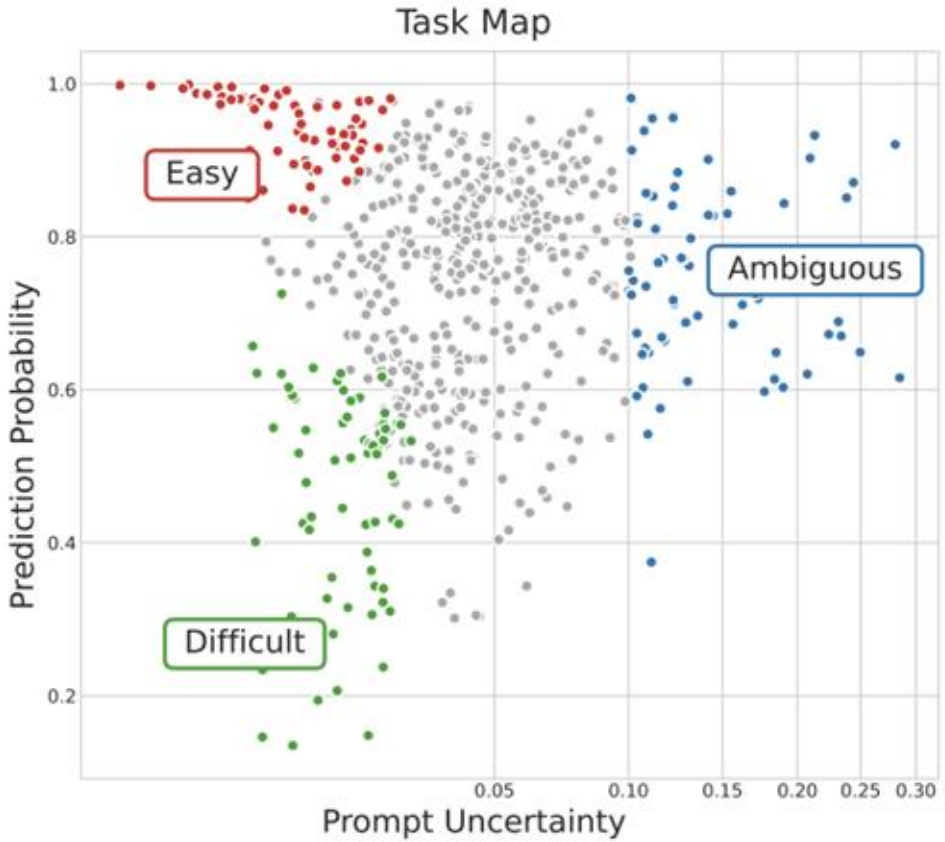
Data Efficiency (1)

- How many fine-tuning data we really need?
- LIMA: Less Is More for Alignment - measure the relative importance of these two stages by training LIMA, a 65B parameter LLaMa language model fine-tuned with the standard supervised loss on only 1,000 carefully curated prompts and responses.



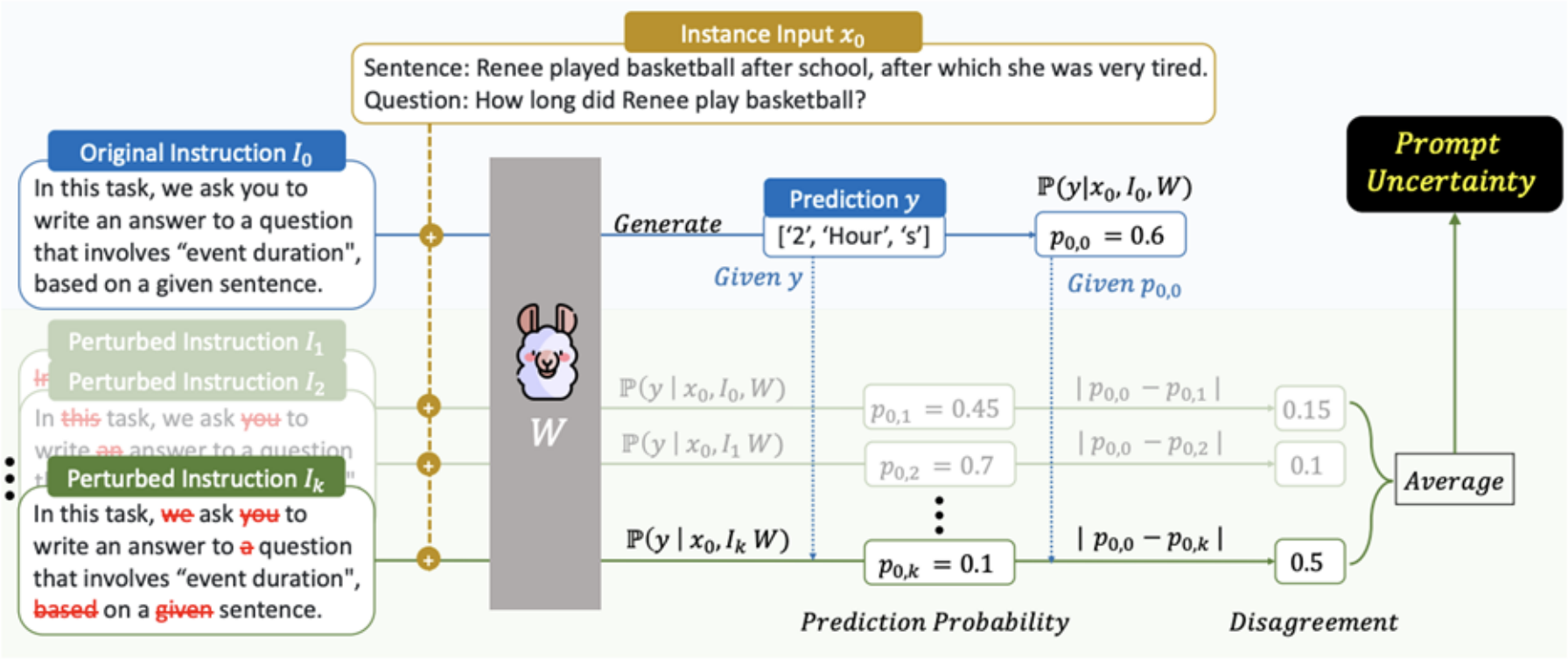
Data Efficiency (2)

- Training a model on a task where the prompt is uncertain will improve its ability to associate the prompt with a specific underlying concept (task), thus achieving better zero-shot performance when the prompt is not seen (easy, ambiguous, difficult).



Data Efficiency (3)

- A proactive instruction tuning method based on Prompt Uncertainty. This method estimates uncertainty by evaluating the degree of inconsistency between the model and the original prediction given complete and perturbed task instructions.



Forms of Data: Alignment Data

- **HelpSteer, HH RLHF, etc.**
  - Often annotated with attributes to help alignment

Name	*Helpfulness-relevant Attributes*	*N conv. (k)*	Mean Length in chars (Std.)	
			Prompt	Response
HELPSTEER	Helpfulness, Correctness, Coherence, Complexity, Verbosity	37.1	2491.8 (1701.7)	497.3 (426.7)
Open Assistant	Quality, Creativity, Humor	59.4	397.5 (620.8)	396.2 (618.8)
HH RLHF	-	337.7	794.4 (706.9)	310.7 (311.4)

Table 1: Overview of Open-source Helpfulness Preference Modeling Datasets

Wang et al ‘23

Forms of Data: Instruction Tuning

- **Natural Instructions**
  - Open dataset
  - Mishra et al, '22
  - 61 tasks, ~200K instructions
    - Note: scale much smaller than pretraining

Example task instances

Instance

- **Input:** Sentence: It's hail crackled across the comm, and Tara spun to retake her seat at the helm.
- **Expected Output:** How long was the storm?

⋮

Instance

- **Input:** Sentence: During breakfast one morning, he seemed lost in thought and ignored his food.
- **Expected Output:** How long was he lost in thoughts?

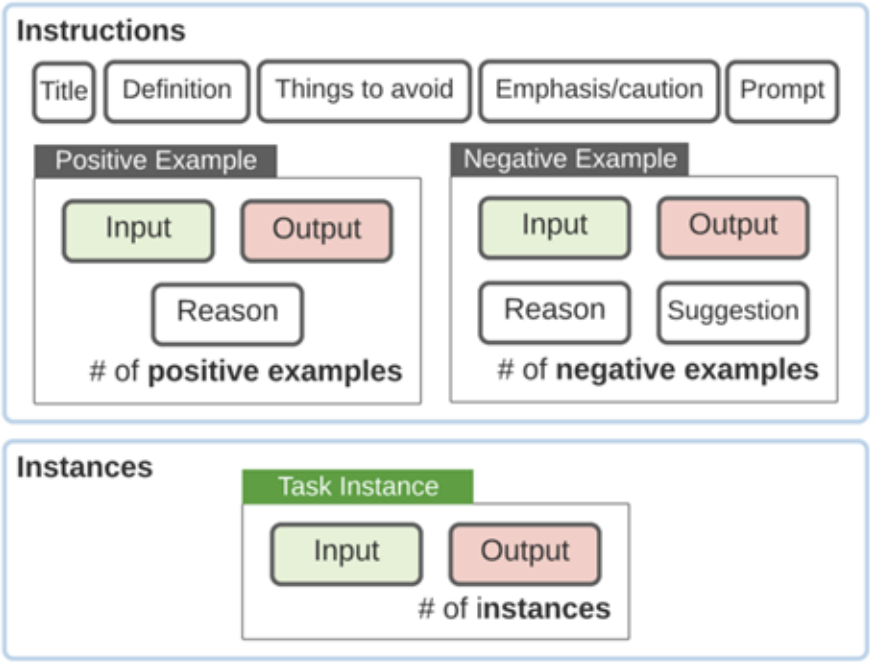
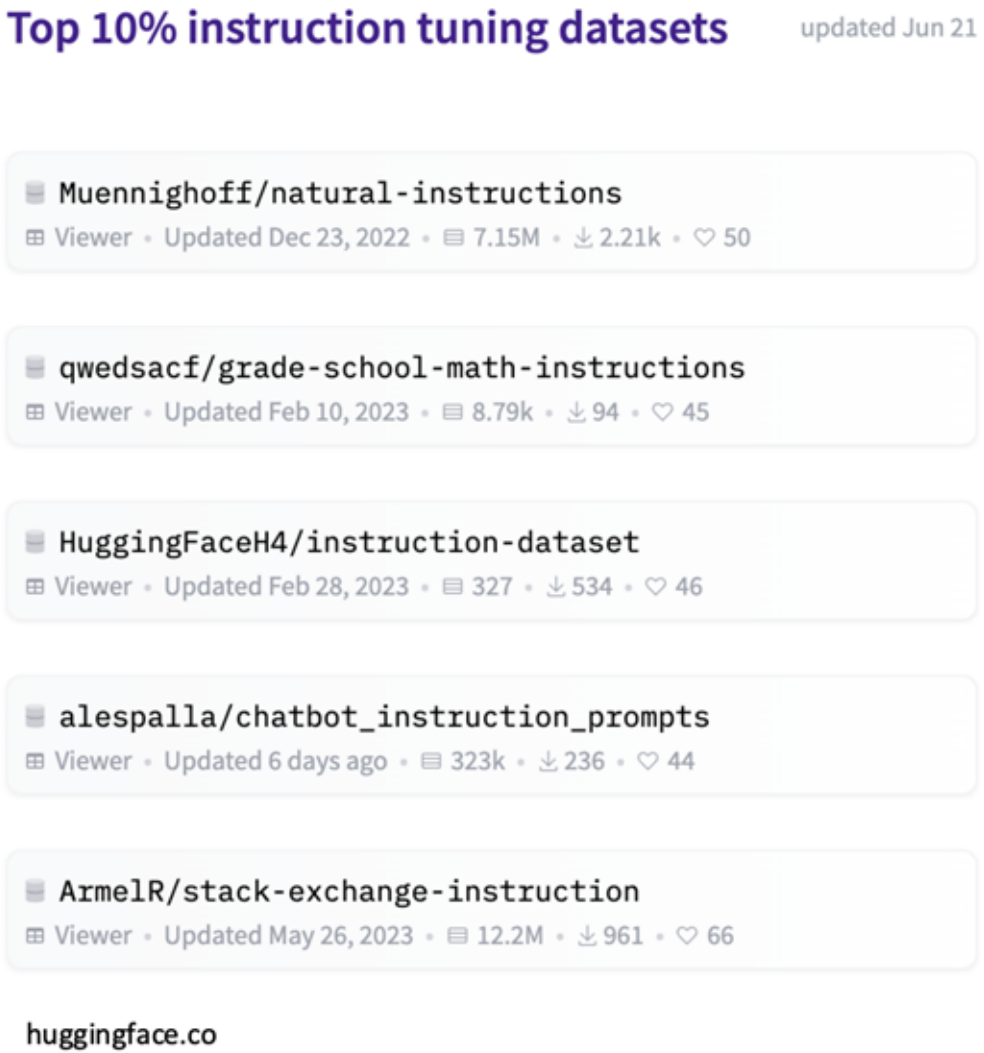


Figure 4: The schema used for representing instruction in NATURAL INSTRUCTIONS (§4.1), shown in plate notation.

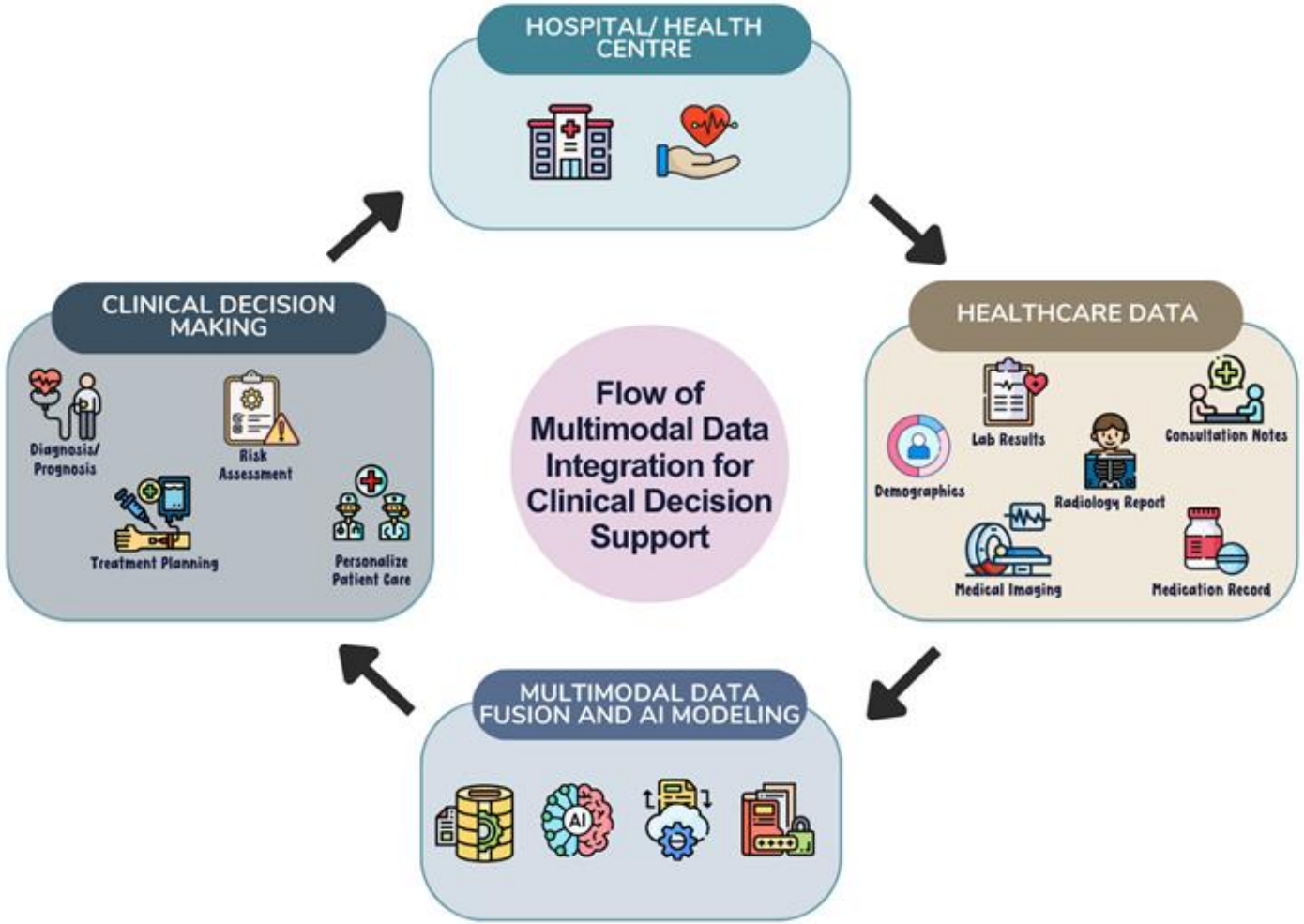
Forms of Data: Instruction Tuning

- **Lots more available,**
  - Hugging face has a great collection
- Pick out ones suitable for your desired instruction-tuned model



Multimodal Data (1)

• Contextual understanding: Allows AI to grasp complex real-world scenarios by seeing, hearing, and reading simultaneously.



Multimodal Data (2)

- Key points of multimodal data: 1) quality of the single modal data; 2) joint content of the two/three modalities; 3) detailed information.

A computer screen with a Windows message about Microsoft license terms.

A can of green beans is sitting on a counter in a kitchen.

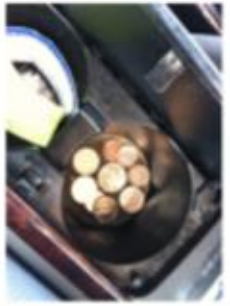
A photo taken from a residential street in front of some homes with a stormy sky above.

A blue sky with fluffy clouds, taken from a car while driving on the highway.

A hand holds up a can of Coors Light in front of an outdoor scene with a dog on a porch.

A digital thermometer resting on a wooden table, showing 38.5 degrees Celsius.

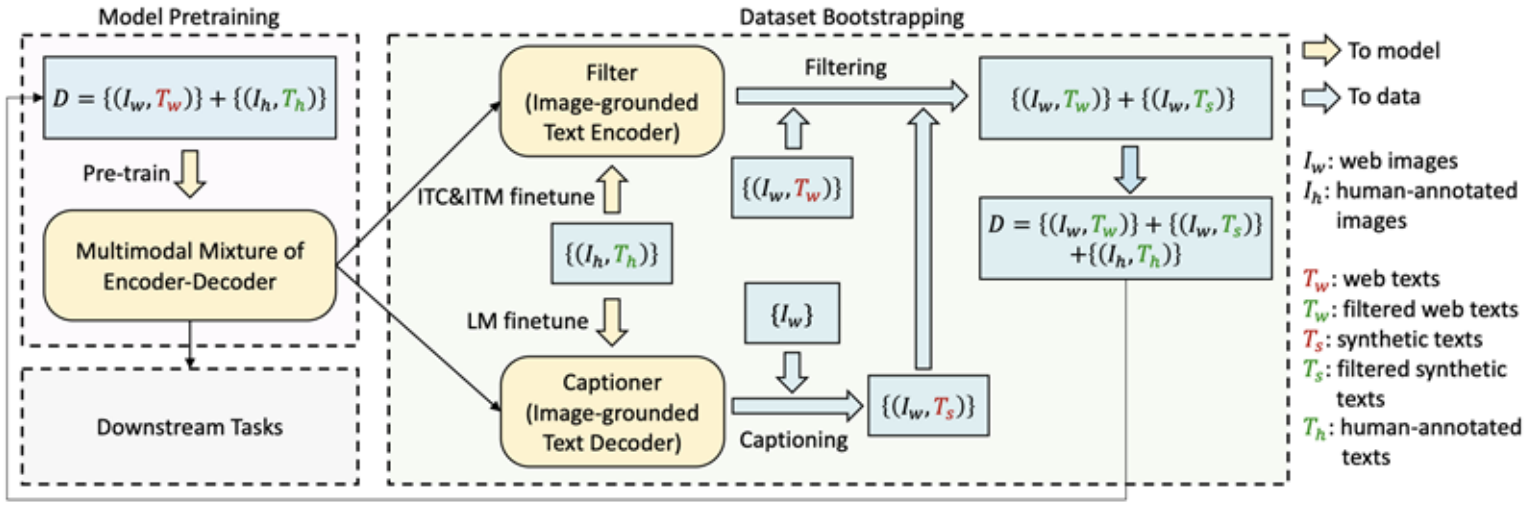
A Winnie The Pooh character high chair with a can of Yoohoo sitting on it in front of a white wall.



A cup holder in a car holding loose change from Canada.

Multimodal Data (3)

- Image: removing blurry images, duplicates, increasing diversity;
- Text: all the methods mentioned in our previous slides



BLIP

T_w: "from bridge near my house"

T_s: "a flock of birds flying over a lake at sunset"

T_w: "in front of a house door in Reichenfels, Austria"

T_s: "a potted plant sitting on top of a pile of rocks"

T_w: "the current castle was built in 1180, replacing a 9th century wooden castle"

T_s: "a large building with a lot of windows on it"

Multimodal Data (4)

- Detailed captions.

LVLM-generated caption: The scene features a large Christmas tree positioned in the middle of a courtyard. The tree is adorned with ornaments and lights, creating a festive atmosphere. A crowd of people is gathered around the tree, enjoying the holiday decorations and the company of others. In total, there are at least 15 people visible in the image, spread out around the tree, some closer to the camera and others further away. The presence of the Christmas tree and the people gathered around it suggest a celebratory and joyful event.

Synthesized caption: In the center of a bustling courtyard, a majestic Christmas tree stands tall, its branches adorned with colorful ornaments and twinkling lights. A group of people, bundled up in warm coats and hats, gather around the tree, chatting and laughing while admiring the festive decorations. In the background, a grand building with a towering dome and a red-bricked facade provides a picturesque backdrop for the holiday scene. Across the street, pedestrians stroll by, bundled up against the cold, while others walk along the sidewalk, taking in the festive atmosphere. The scene exudes a sense of joy and celebration, capturing the spirit of the holiday season.

LVLM-generated caption: The image depicts a bustling street scene with several shops and a marketplace. People are walking around and interacting with each other, creating a lively atmosphere. Some individuals are standing in front of a building, while others are scattered throughout the scene. In the background, a large tree is present, adding a touch of greenery to the environment. The street is lined with various signs and banners, providing information and advertisements for the shops and market. The street is wide enough to accommodate the pedestrian traffic, and the overall scene gives off a feeling of a busy market day.

Synthesized caption: In the bustling street scene, a man stands outside a building with a street sign, while a person can be seen standing in front of a doorway. A man sits on a bench in front of a shop sign, adding to the lively atmosphere. The sidewalk is lined with signs, including a sign that says "Quizuara". In the background, a tall tree with no leaves casts a shadow, while power lines stretch across the sky, barely visible. The brick road is wide enough to accommodate the pedestrian traffic, creating a busy market day feel.

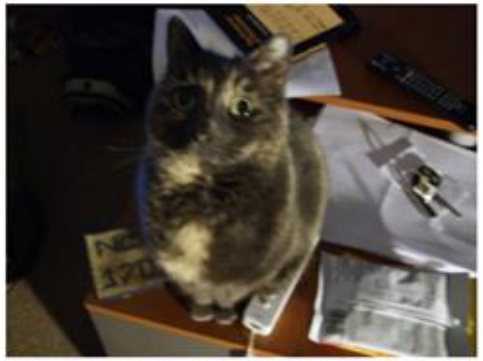
Multimodal Data (5)

- Image captioning hallucination.

Image Model predictions:
bowl, broccoli, carrot, dining table

Language Model predictions for the last word:
fork, spoon, bowl

Generated caption: A plate of food with broccoli and a ***fork***.



TopDown: A cat sitting on top of a ***laptop computer***.
NBT: A cat sitting on a table next to a ***computer***.

TopDown: A brown dog sitting on top of a ***chair***.
NBT: A brown and white dog sitting under an ***umbrella***.

TopDown: Aa man and a woman are playing with a ***frisbee***.
NBT: A man riding a skateboard down a street.

TopDown: A man standing on a beach holding a ***surfboard***.
NBT: A man standing on top of a sandy beach.

Multimodal Data Resource

- CC3M

by Joi Ito

the trail climbs steadily uphill most of the way.

by Danail Nachev

the stars in the night sky.

by Justin Higuchi

musical artist performs on stage during festival.

by Viaggio Routard

popular food market showing the traditional foods from the country.

Creating Conceptual Captions

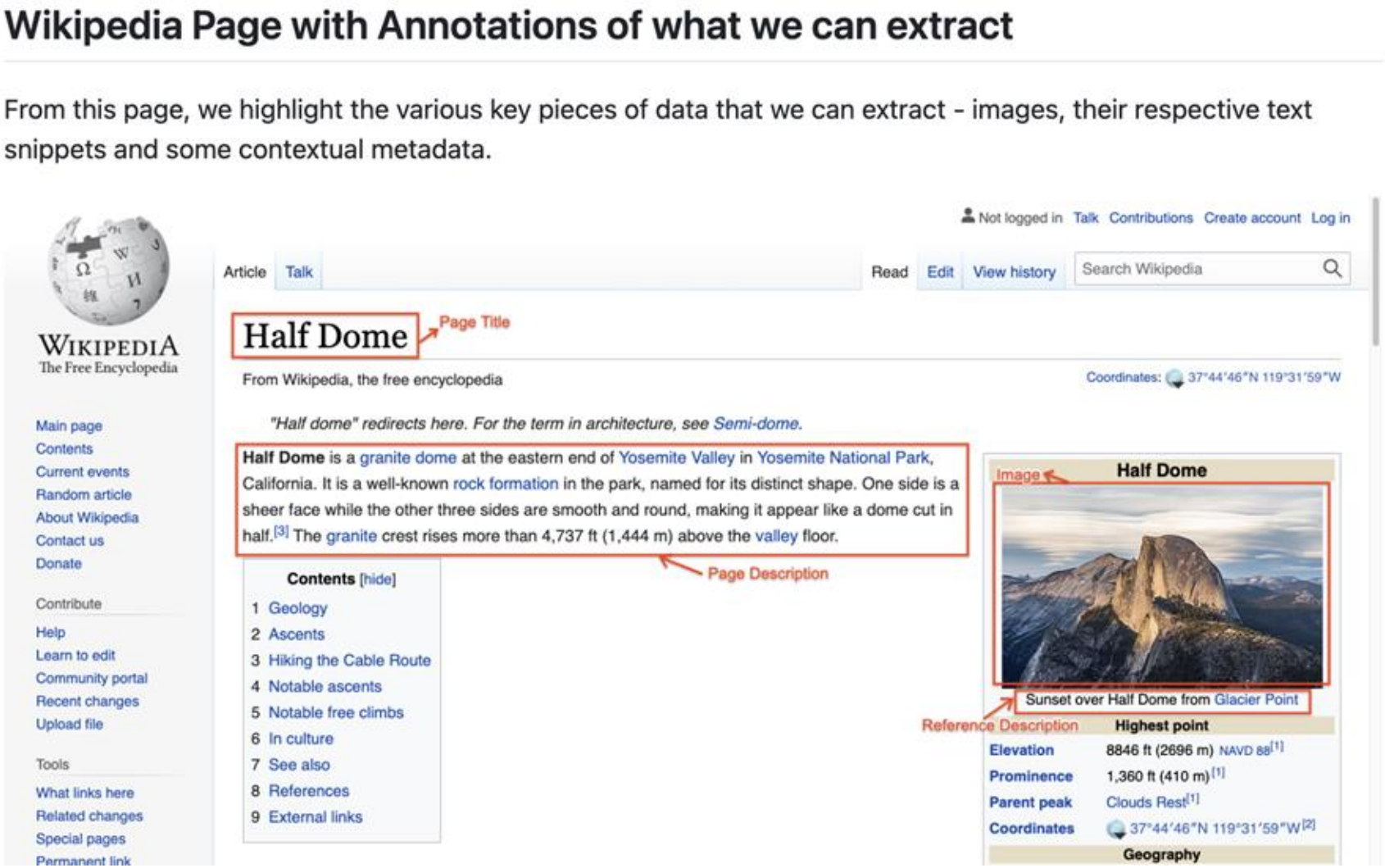
For Conceptual Captions, we developed a fully automatic pipeline that extracts, filters, and transforms candidate image/caption pairs, with the goal of achieving a balance of cleanliness, informativeness, fluency, and learnability of the resulting captions. Because no human annotators are

Dataset Stats

The resulting dataset (version 1.1) has been split into Training, Validation, and Test splits. The Training split consists of 3,318,333 image-URL/caption pairs, with a total number of 51,201 total token types in the captions (i.e., total vocabulary). The average number of tokens per captions is 10.3 (standard

Multimodal Data Resource

- Web Image-Text-1B



Multimodal Data Resource

- OBELICS

Image-Text Pairs

Tottenham vs Chelsea Live Streaming

Tottenham Spurs vs Chelsea Live Streaming

Multimodal Document

The match between Tottenham Spurs vs Chelsea will kick off from 16:30 at Tottenham Hotspur Stadium, London.

The derby had been played 54 times and the Blues have dominated the Spurs. Out of 54 matches played, Chelsea has won 28 times and Spurs had only won 7 times. The remaining 19 matches had ended in draw.

However, in recent 5 meetings, Spurs had won 3 times where Chelsea had won the other two times. …

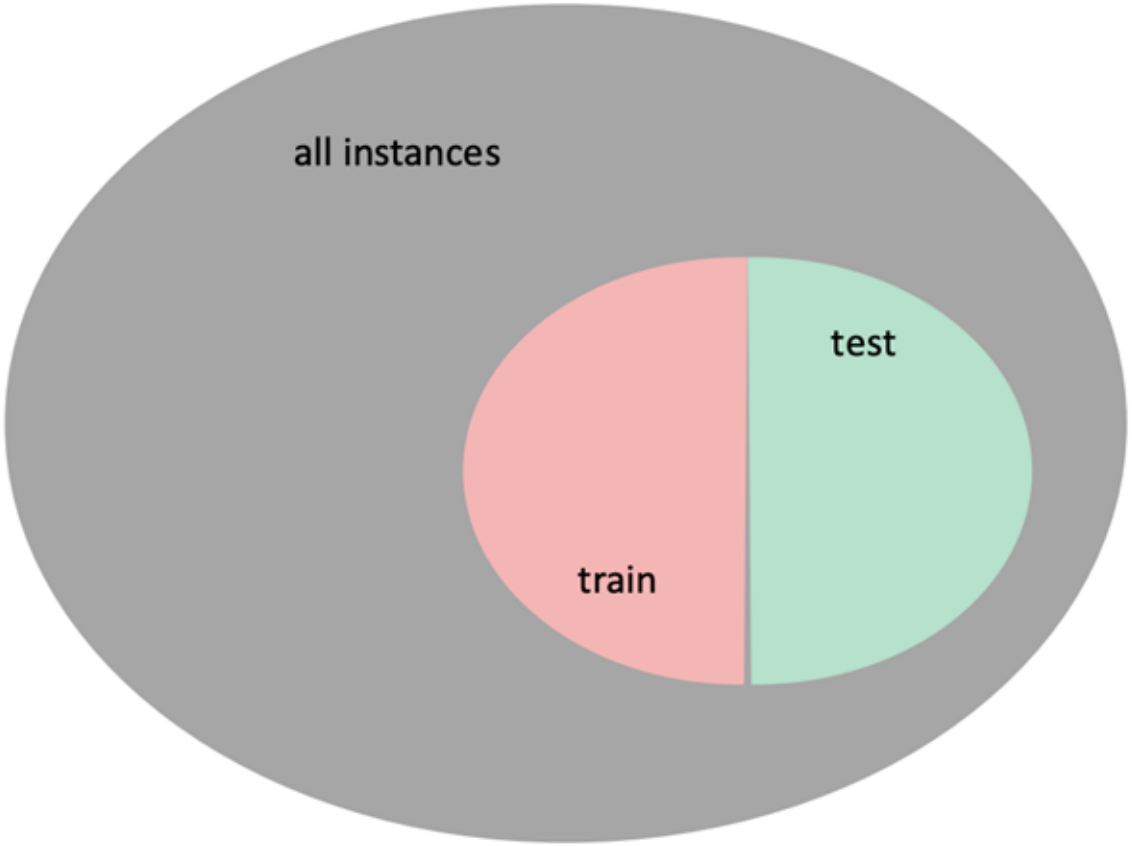
Evaluation (1)

Traditional approach in ML:

- Measure accuracy or a related metric on a test set
- Or perform cross-validation, etc.

- Can switch from accuracy to other metrics: AUC-ROC, F1 non-scalar metrics like the confusion matrix, etc.

Could still do some of these…



Evaluation (2)

For large language models, a bit more complex

- Far more general capabilities
- Space of outputs much larger than multiclass classification!
  - Many answers might be right!

- What do we need for an evaluation system? Some pieces:
  - Dataset
  - Metrics
  - Mechanism to compute metrics on model

Evaluation - HumanEval

Chen et al ’21 introduced **Codex**

- Essentially a fine-tuned version of GPT3 for code
- How to evaluate?
  - Output is now code---lots of ways to write “good” code

- **Example**:

```
def incr_list(l: list):
    """Return list with elements incremented by 1.
    >>> incr_list([1, 2, 3])
    [2, 3, 4]
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])
    [6, 4, 6, 3, 4, 4, 10, 1, 124]
    """
    return [i + 1 for i in l]
```

Evaluation - HumanEval

Chen et al ’21 introduced **Codex**

- What do we need for an evaluation system?
- **Dataset**: “a set of 164 handwritten programming problems”
  - Each problem: definition, some metadata, variable # of test cases
  - “Programming tasks … assess language comprehension, reasoning, algorithms, and simple mathematics”

```
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

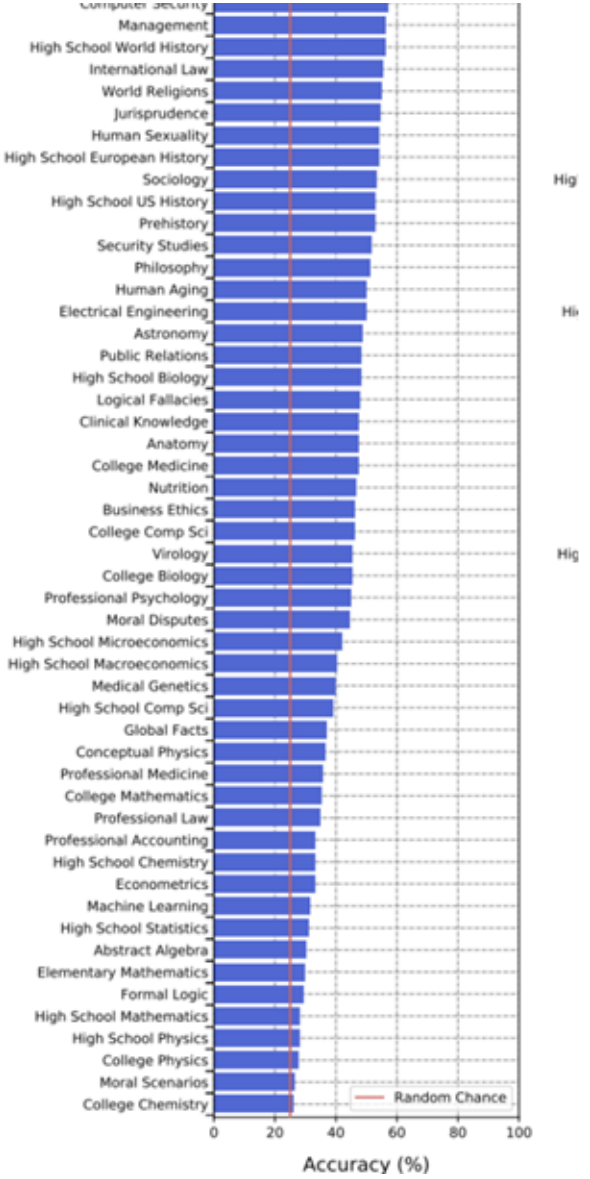
    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

Evaluation - MMLU

Hendrycks et al ’21 MMLU

- “Measuring Massive Multitask Language”
- **Idea**: measure model knowledge
  - 0-shot or few-shot
  - Do this across many different areas: 57 total across high school / college settings
  - 15908 total questions

- Note: models are quite good at MMLU now!
  - But GPT3 still struggled on certain areas back then
  - Still in use!



Evaluation - MMLU

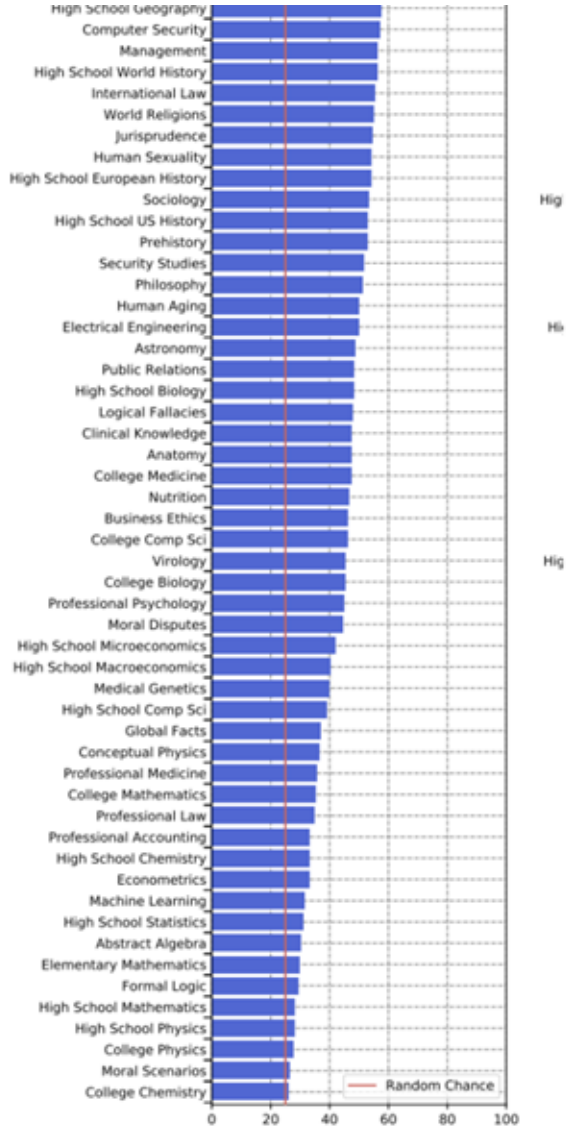
Hendrycks et al ’21 MMLU

- **Dataset**: 15908 Qs from 57 areas
- All multiple choice with 4 options
- Validation/test split: 1540/14079 Qs
- Example:

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points
(B) circle
(C) half-line
(D) line

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.



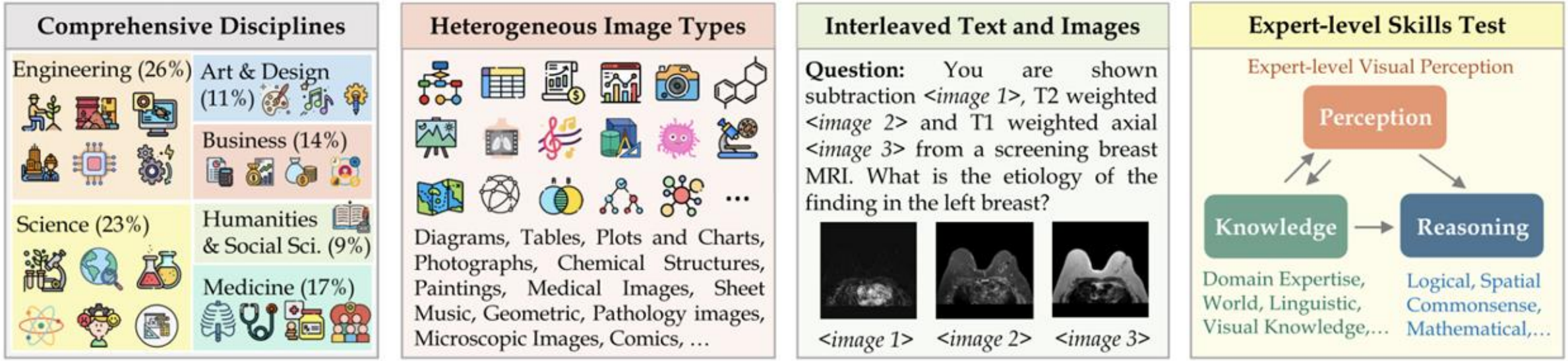
Evaluation - MMLU-Pro

Improvements:

- MMLU has 4 multiple choice answers, MMLU-Pro has 10
  - I.e., more possible “distractors”
- MMLU predates chain-of-thought, so most questions are not affected by CoT. MMLU-Pro has more “reasoning” type questions
- Expert reviews for questions (question noise a major issue)
- More flexibility in answering,
  - “use the regular expression ‘answer is \(?\([A-J]\)?\)’”

Evaluation - MMMU

- MMMU-Multi-discipline Multimodal Understanding and Reasoning



Overview of the MMMU dataset. MMMU presents four challenges: 1) **comprehensiveness**: 11.5K college-level problems across six broad disciplines and 30 college subjects; 2) highly **heterogeneous** image types; 3) **interleaved** text and images; 4) **expert-level** perception and reasoning rooted in deep subject knowledge.

Thanks

Q&A

Prof. Yu Cheng
chengyu@cse.cuhk.edu.hk