PAPER

Datasets for Large Language Models: A Comprehensive Survey

Without high-quality datasets as the foundation, it is challenging to grow the tree of LLMs with flourishing branches and leaves.

The composition and quality of these datasets profoundly influence the performance of LLMs.

What the paper talks about?

- summarizes existing representative datasets across five dimensions: pre-training corpora, instruction fine-tuning datasets, preference datasets, evaluation datasets, and traditional NLP datasets

pre-training corpora

- General pre-training corpora categorized by data types and domain-specific pre-training corpora categorized by domains.
- The pre-training corpora are large collections of text data used during the pre-training process of LLMs
- In the pre-training phase, LLMs learn extensive knowledge from massive amounts of unlabeled text data (stored in its model parameters)
- The pre-training corpora have various types of text data: webpages, academic materials, books, while also accommodating relevant texts from diverse domains: legal documents, annual financial reports, medical textbooks, and other domain-specific data.
- Domains involved in the pre-training corpora can be divided in 2 types: general pre-training corpora and domain-specific pre-training corpora

---

general pre-training corpora

- large-scale text data mixtures from different domains and topics
- objective: provide universal language knowledge and data resources for NLP tasks

domain-specific pre-training corpora

- contains relevant data for specific domains or topics
- objective: furnish LLMs with specialized knowledge

- The pre-training corpora influence the direction of pre-training and the potential of models in the future.

General Pre-training Corpora

Language Texts

The language text data mainly consists of two parts.

The first part is electronic text data constructed based on widely sourced written and spoken language, typically in the form of large corpora for a specific language. The full name of ANC is the American National Corpus. The content primarily includes various written and spoken materials in American English. **The second edition of the corpus has a scale of 22M words, making it highly suitable for models to learn language.** (Espanhol - ATENÇÃO modelos pre-treinados na 2ª edição do ANC)

The second part is electronic text data constructed based on relevant written materials in various fields or topics.

Books

- Improves their ability to capture human language features while learning more profound language knowledge and contextual information.

The book data primarily possesses the following characteristics.

- **Breadth** - It typically covers a wide range of subjects and topics, including novels, biographies, textbooks, and more.
- **High Quality** - Books are usually authored by professionals, undergo editing and proofreading, resulting in more accurate grammar and spelling with less noise.
- **Lengthy Text** - Longer texts and complex sentence structures provide additional contextual information.
- **Language and Culture** - Books often contain rich language features such as professional terminology, colloquialisms, and idioms, reflecting diverse cultural backgrounds.

Encyclopedia

The most common encyclopedia corpus is Wikipedia.

Benchmark (shots)	GPT-3.5	GPT-4	PaLM	PaLM-2-L	LLAMA 2
MMLU (5-shot)	70.0	**86.4**	69.3	78.3	68.9
TriviaQA (1-shot)	–	–	81.4	**86.1**	85.0
Natural Questions (1-shot)	–	–	29.3	**37.5**	33.0
GSM8K (8-shot)	57.1	**92.0**	56.5	80.7	56.8
HumanEval (0-shot)	48.1	**67.0**	26.2	–	29.9
BIG-Bench Hard (3-shot)	–	–	52.3	**65.7**	51.2

Multi-category Corpora

- Multi-category corpora contain two or more types of data, which is beneficial for enhancing the generalization capabilities of LLMs.

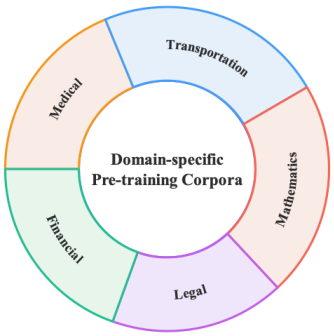

Fig. 6 Domain categories of the domain-specific pre-training corpora

Domain-specific pre-training corpora

- The type of corpus is typically employed in the incremental pre-training phase of LLMs.
- After training a base model on a general pre-training corpus, if the **model needs to be applied to downstream tasks in a particular domain**, domain-specific pre-training corpora can be further utilized to incrementally pre-train the model.

Domain-specific Pre-training Corpora

Attention: Training with an excess of data from a particular domain can impact the generalization ability of LLMs in other domains.

Preprocessing of Pre-training Data

- The collected data needs to undergo a preprocessing pipeline to enhance data quality and standardization while reducing harmful and sensitive content.
- Data preprocessing generally consists of five steps: (1) Data Collection. (2) Data Filtering. (3) Data Deduplication. (4) Data Standardization. (5) Data Review.

(1) Data Collection

A comprehensive data collection phase involves:

Define Data Requirements
Define data requirements, including data types, language, domain, sources, quality standards helps **determine the scope and objectives of data collection**.

Select Data Source
Selecting appropriate data sources can include: websites, books, academic papers, and other resources.

We can use filters to, for example, exclude low-quality websites.

Develop Collection Strategy
The collection strategy encompasses the time span, scale, frequency, and methods of data collection, **facilitating the acquisition of diverse and real-time data**.

Data Crawling (Rastreamento de Dados) and Collection
Collect text data from the selected data sources according to the predefined collection strategy.

Ensure compliance with **legal regulations and the relevant agreements and policies of the websites during the crawling process**.

Data Extraction and Parsing
Extract textual components from raw data, **enabling accurate parsing and separation of text**.

Encoding Detection
Encoding detection tools -> ensuring that text is stored in the correct encoding format.

Incorrect encoding may lead to garbled characters or data corruption.

Language Detection
Language detection tools -> identify the language of the text, enabling the segmentation of data into subsets based on different languages, **selecting only the required language texts**.

Data Backup
Advice : periodically back up the collected data to **prevent data loss and damage**.

Privacy and Legal Compliance
Data privacy laws and regulations, obtain necessary permissions, and protect personal and sensitive information in the data.

Maintenance and Updates
Regularly maintain the data collection system to **ensure the continuous updating of data**.

Consider replacing with new data sources and collection strategies as needed.

(2) Data Filtering

Data filtering is the process of screening and cleaning the data obtained during the data collection stage, with the primary goal of **improving data quality**. Can be accomplish through model-based methods or heuristic-based methods.

Model-based methods

High-quality pre-training corpora can be used as **positive** samples, with the contaminated text to be filtered as **negative** samples, to train classifiers for filtering.

Example : WanJuanText-1.0

1 - They train content safety models for both Chinese and English content to filter potential harmful data related to topics like obscenity, violence, and gambling.

2 - They train data quality models for both Chinese and English to address low-quality contents such as advertising and random data in webpages, thereby reducing the prevalence.

Heuristic-based models

Filtering can be conducted at both the document level and sentence level.

Document level : employing heuristic rules to delete entire documents in the corpus that do not meet the requirements.

Sentence level : heuristic rules to delete specific sentences within a document that do not meet the criteria.

This heuristic rules are manually defined.

At the document level, most corpora undergo language filtering to exclude unwanted documents.This step can also be completed during the **language detection phase of data**

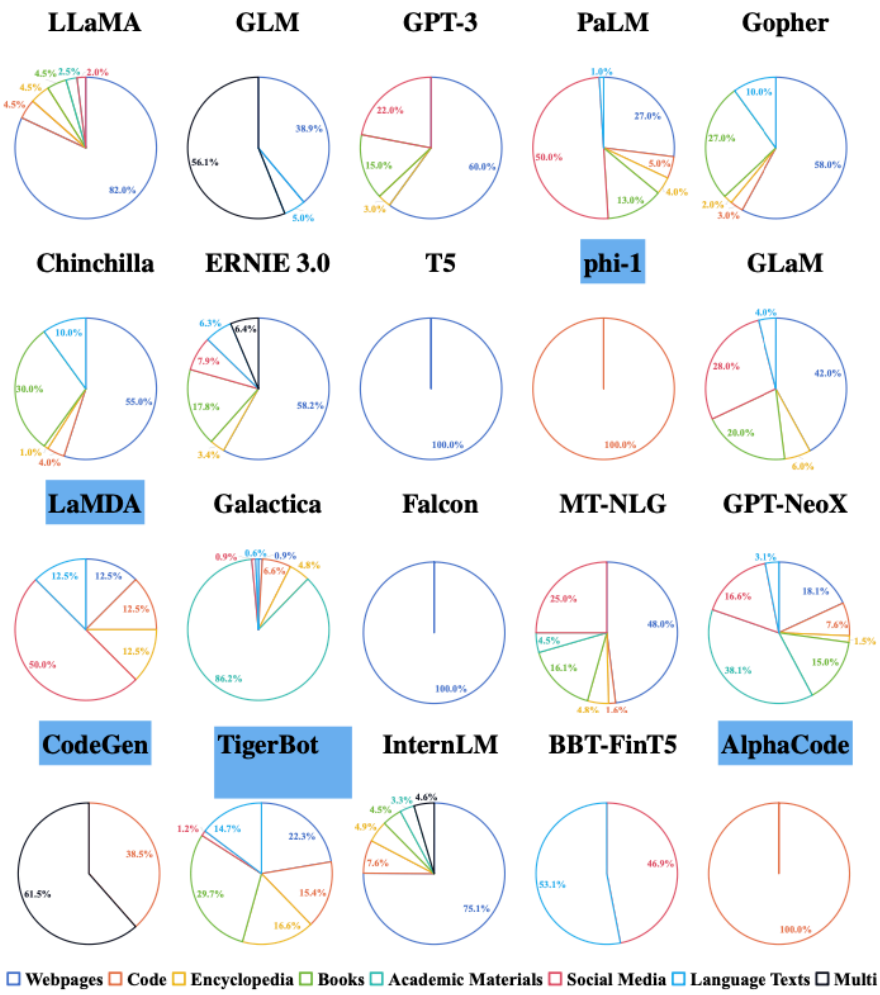

Fig. 8 The distribution of data types in pre-training corpora used by different LLMs. Each pie chart displays the name of an LLM at the top, with different colors representing various data types

collection.

At the sentence level, corresponding heuristic rules are set to selectively remove sentences that are not necessary to retain in the corpus.

(3) Data Deduplication

Data deduplication involves removing duplicate or highly similar texts in a corpus.

Typical deduplication methods:

1. TF-IDF (Term Frequency-Inverse Document Frequency) Soft Deduping
2. MinHash
3. SimHash

(4) Data Standardization

Data standardization involves the normalization and transformation of text data to make it more manageable and comprehensible during the model training process.

Consists of 4 steps:
- Sentence Splitting
- Simplified Chinese
- Spelling Correction
- Remove Stop Words

(5) Data Review

1. **Documenting the previous preprocessing steps and methods** for future reference and review.
2. A manual **review** is done to **check if the data processing meets the expected standards** -> Any issues identified during this review are then provided as feedback to steps 1 through 4 (Can also be done at the end of every step from 1 to 4).

instruction fine-tuning datasets

The instruction fine-tuning datasets consists of a series of text pairs comprising “instruction inputs” and “answer outputs.”

Instruction inputs : **requests** made by **humans** to the model (classification, sum- marization, paraphrasing)

Answer outputs : **responses** generated by the **model** following the instruction, looking for the human expectations

There are 4 ways to construct instruction fine-tuning datasets:

1. manual creation
2. model generation, for example, using the Self-Instruct method
3. collection and improvement of existing open-source datasets
4. a combination of the three aforementioned methods.

The instruction fine-tuning datasets are used to further fine-tune pre-trained LLMs.

The instruction fine-tuning datasets can be divided into two main categories: **general instruction fine-tuning datasets** and **domain-specific instruction fine-tuning datasets**.

general instruction fine-tuning datasets
aiming to enhance the models’ performance across a wide range of tasks

domain-specific instruction fine-tuning datasets
the instructions are specifically designed for particular domains

Instructions are broadly grouped into 15 classes: Reasoning, Math, Brainstorming, Closed QA, Open QA, Code, Extraction, Generation, Rewrite, Summarization, Translation, Role-playing, Social Norms, and Others

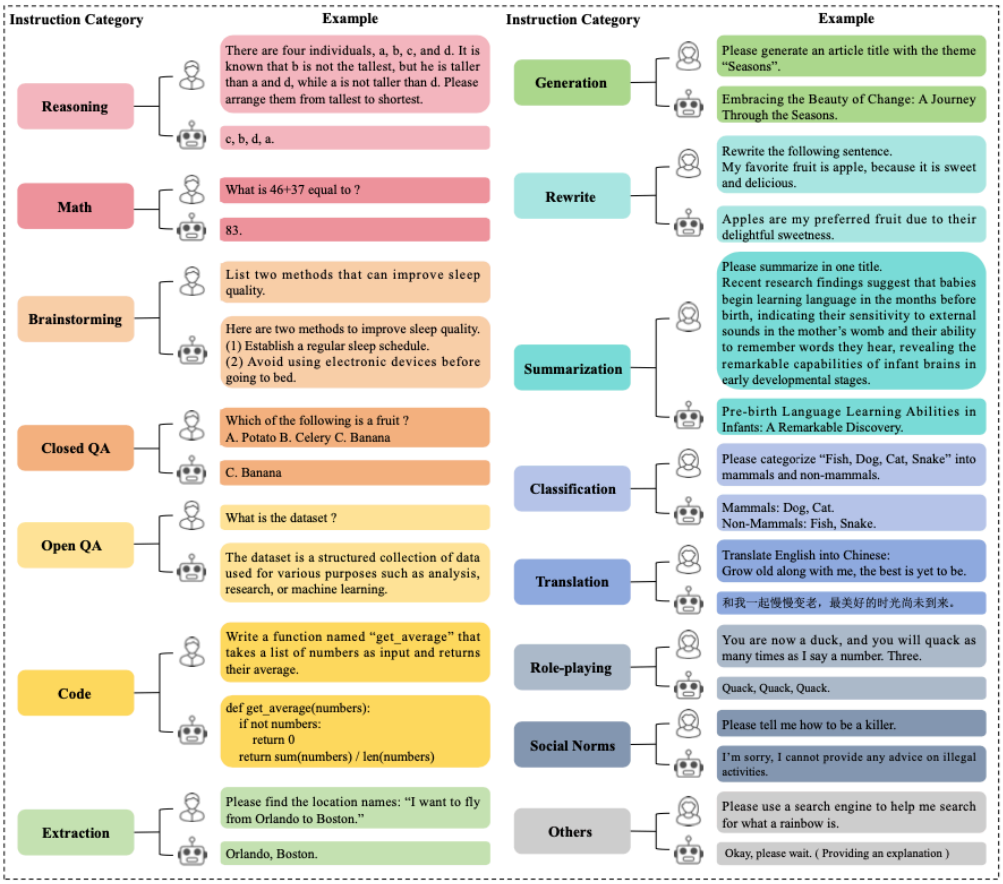

Fig. 10 Summary of instruction categories, which are categorized into 15 groups

We are going to work with: Open QA

For Open QA instructions, questions do not come with options, and answers cannot be directly found within the question. One must rely on their own knowledge base to formulate a response. These questions can include com- mon knowledge queries with standard answers or open-ended inquiries without predefined solutions.

General Instruction Fine-tuning Datasets

- one or more instruction categories
- no domain restrictions
- Why? Aiming to enhance the instruction-following capability of LLMs in general tasks.

General instruction fine-tuning datasets are categorized into four main types based on their construction methods: Human Generated Datasets, Model Constructed Datasets, Collection and Improvement of Existing Datasets, and Datasets Created with Multiple Methods.

Nao aprofundei!

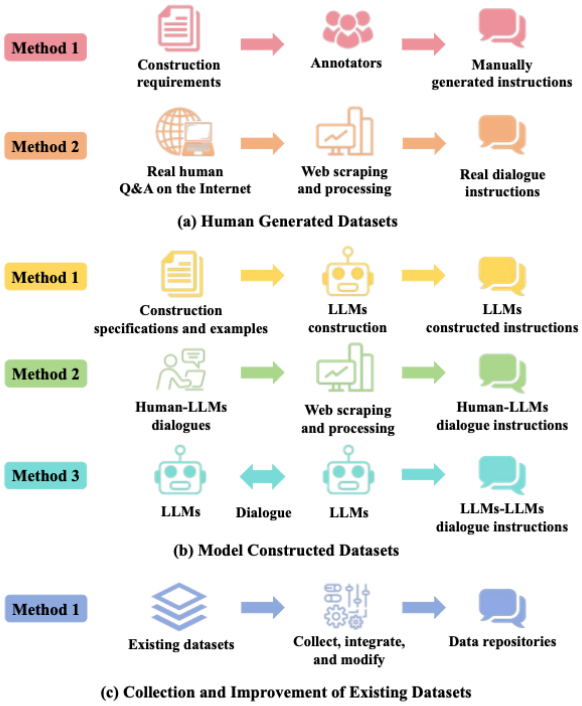

Fig. 12 Different approaches to instruction construction

Domain-specific Instruction Fine-tuning Datasets

The domain-specific instruction fine-tuning datasets are constructed for a particular domain by formulating instructions that encapsulate knowledge and task types closely related to that domain.

preference datasets

The feedback information in preference datasets is often manifested through voting, sorting, scoring, or other forms of comparison.

Both RLHF and RLAIF (Reinforcement Learning from AI Feedback) employ reinforcement learning methods to optimize models using feedback signals.

Preference Evaluation Methods

Each method can be conducted by humans or aligned high-quality LLMs. Both Humans and LLMs have their advantages and disadvantages.