Domain Composition (Sec. 2.1)
- Web — Raw text
- Wiki — Raw text
- ......
- Code — Raw text

Quantity Control (Sec. 2.2)

Scaling laws
- Kaplan's,
- Chinchilla

Repetition
- multi-epoch,
- repeating on selected data

Quality Control (Sec. 2.3)

Quality Filtering (Sec. 2.3.1) → Deduplication (Sec. 2.3.2) → Toxicity Filtering (Sec. 2.3.3)

Quality Filtering (Sec. 2.3.1)
- Heuristics,
- Classifiers,
- Metric thresholding,
- Clustering

Deduplication (Sec. 2.3.2)
- N-gram-and-hashing,
- Line/document-level,
- Neural models,
- Semantic clustering

Toxicity Filtering (Sec. 2.3.3)
- Heuristics,
- Classifiers

Diversity & Age
- Diversity-difficulty balancing data selection,
- Temporal shift between evaluation and pretraining data

Pretraining Dataset

(a) Data management pipeline in the pretraining stage of LLMs

Task Composition (Sec. 3.1)
- Task 1 — Raw instructions
- Task 2 — Raw instructions
- ......
- Task N — Raw instructions

Quality Control (Sec. 3.2)

Quality (Sec. 3.2.1)
- Heuristic/Model-based indicators,
- LLMs as quality judges

Diversity (Sec. 3.2.2)
- Quantitative measure,
- Similarity/Distance-based filtering,
- Augmentation with active searching

Complexity (Sec. 3.2.3)
- Quantitative measure,
- Incremental augmentation

Quantity Control (Sec. 3.3)
- Scaling up v.s. Scaling down,
- Scaling patterns for different abilities

Instruction Dataset

Dynamic Data-Efficient Learning (Sec. 3.4)
- Learning Affects Data
- Data Affects Learning

LLM Fine-tuning

(b) Data management pipeline in the supervised fine-tuning stage of LLMs