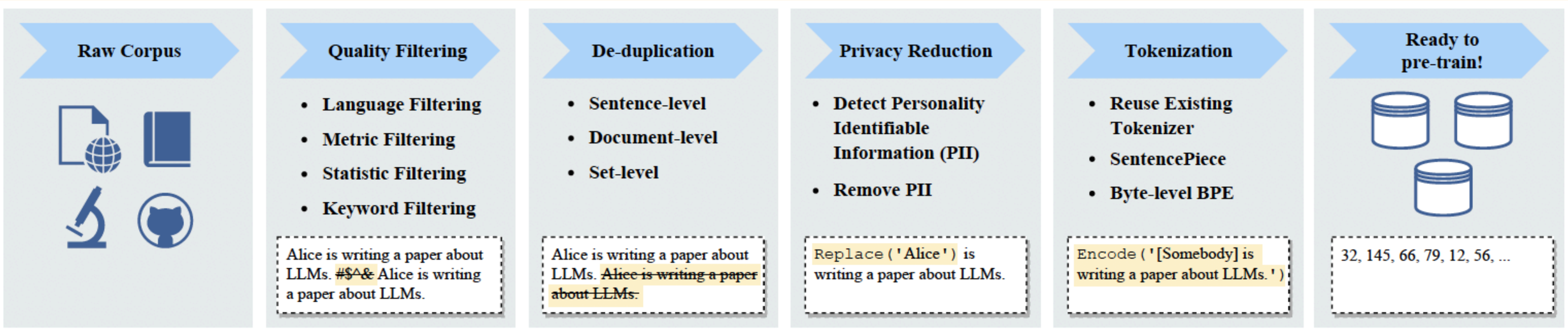

Fig. 6: An illustration of a typical data preprocessing pipeline for pre-training large language models.

LLMに対する事前学習データの効果

- ソースの混合
  - ❖ 異なるソースを混合する場合、できるだけ多くの高品質なデータソースを含めることが推奨される
  - ❖ 下流のタスクにおけるLLMのパフォーマンスにも影響する可能性が高いため慎重に設定する必要がある
- 事前学習データの量
  - ❖ LLMのパラメータスケールが大きくなるにつれて、訓練に必要なデータ量が増加する
- 事前学習データの質
  - ❖ 適切にフィルタリングしたクリーンなデータで訓練すると性能は向上する
  - ❖ 重複データはLLMの文脈からのコピー能力を低下させる (In-context learningを阻害する)