Language
Filtering

Deduplication
by URL

Quality* Filters
C4 (subset) + Gopher rules

Content Filters
Toxic content, PII

Deduplication
on text overlap

Decontamination
against eval set