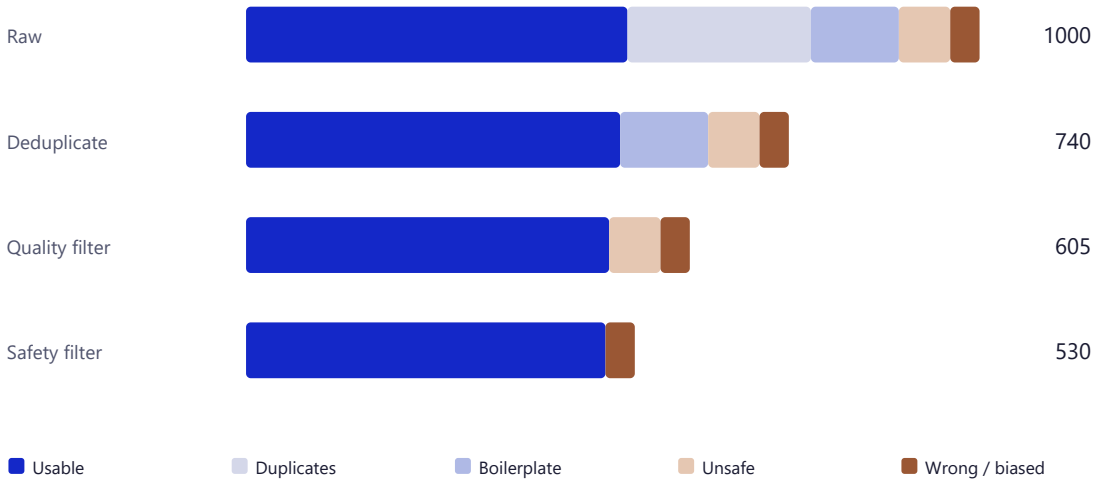
CHAPTER 06 / CURATION

Data is a design decision.

Follow what survives a training-data pipeline.

Follow what survives a training-data pipeline.



40 wrong / biased documents survive every filter. No stage checks truth.

Invented corpus and independent removal rules. No stage here checks truth.

Coverage

Sources and cutoff bound what could have been learned. A corpus is not a verified lookup table.

Curation

Filters remove unwanted text and can also discard usable documents.

Mixing

Sampling weights change emphasis without adding documents.

↗ **Sources and cutoff bound what could have been learned. A corpus is not a verified lookup table.**

PUT IT TO WORK

Toy corpus: 1,000 documents become 530 after three filters. Forty wrong or biased documents survive every stage.

CLOSE THE PAGE. EXPLAIN IT.

Which fact your task needs is likely missing from training, and how will you supply or verify it?