

The estimate looked rigorous — the critic caught the invented number

The first attempt named capture-recapture, then hard-coded a guess: "each source holds ~100 docs" — the reality is 451-1498 docs.
A scalar score only says the number is wrong; the HDS6 process critic diagnosed WHY, and drove the same model to measure them.

TRAJECTORY 1

baseline · claude-sonnet-5

same step in both runs

HDS6 PROCESS CRITIC

grades trajectory 1 · its note is fed into trajectory 2

first attempt

improved

critic's finding

the fix

TRAJECTORY 2

same model + repair note

Assemble a pretraining corpus
sample only



How many unique, useful docs can we get?

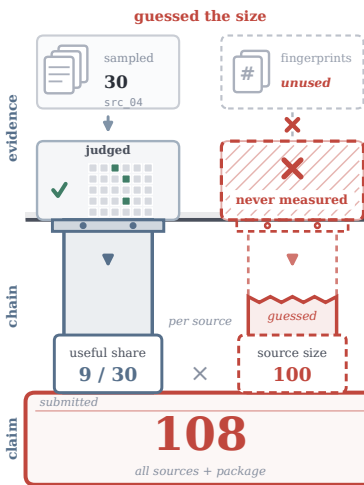
Sample the sources



Judge samples



Preview packages



estimation score



P1

Read the task

Estimate how many UNIQUE high-value docs the sources hold, on a tight budget. Scaling up a small sample UNDER-counts — use capture-recapture on the cross-source overlap.

P2

Sample the sources

It can't read the sources in full — it draws small metered samples from each source.

P3

Judge the samples

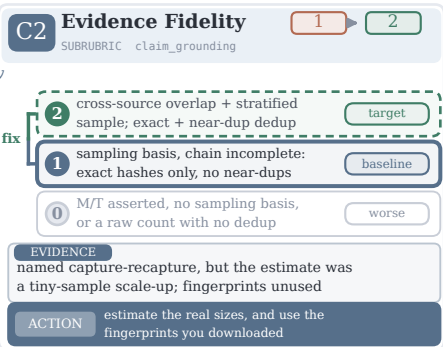
An LLM-judge marks each sampled doc high-value or spam, so a source's quality is read off its sample.

P4

Preview packages

Two purchasable packages are previewed — the preview lists only the content hashes (fingerprints) of the docs, which is how repeats get spotted.

P5 · Estimate each source's size — critic's finding



P6

Estimation

truth ~1505 unique docs · 108 is 93% off, 2471 is 64% off
score = $0.7 \times (1 - \text{error})$, averaged over docs and tokens

Assemble a pretraining corpus
sample only



How many unique, useful docs can we get?

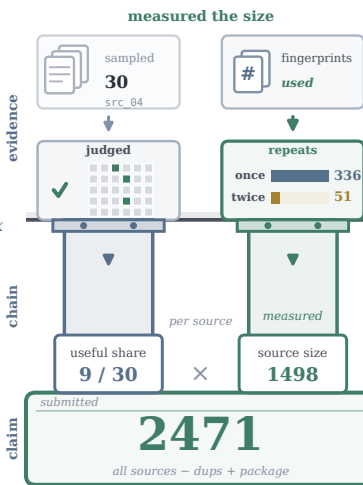
Sample the sources



Judge samples



Preview packages



estimation score

