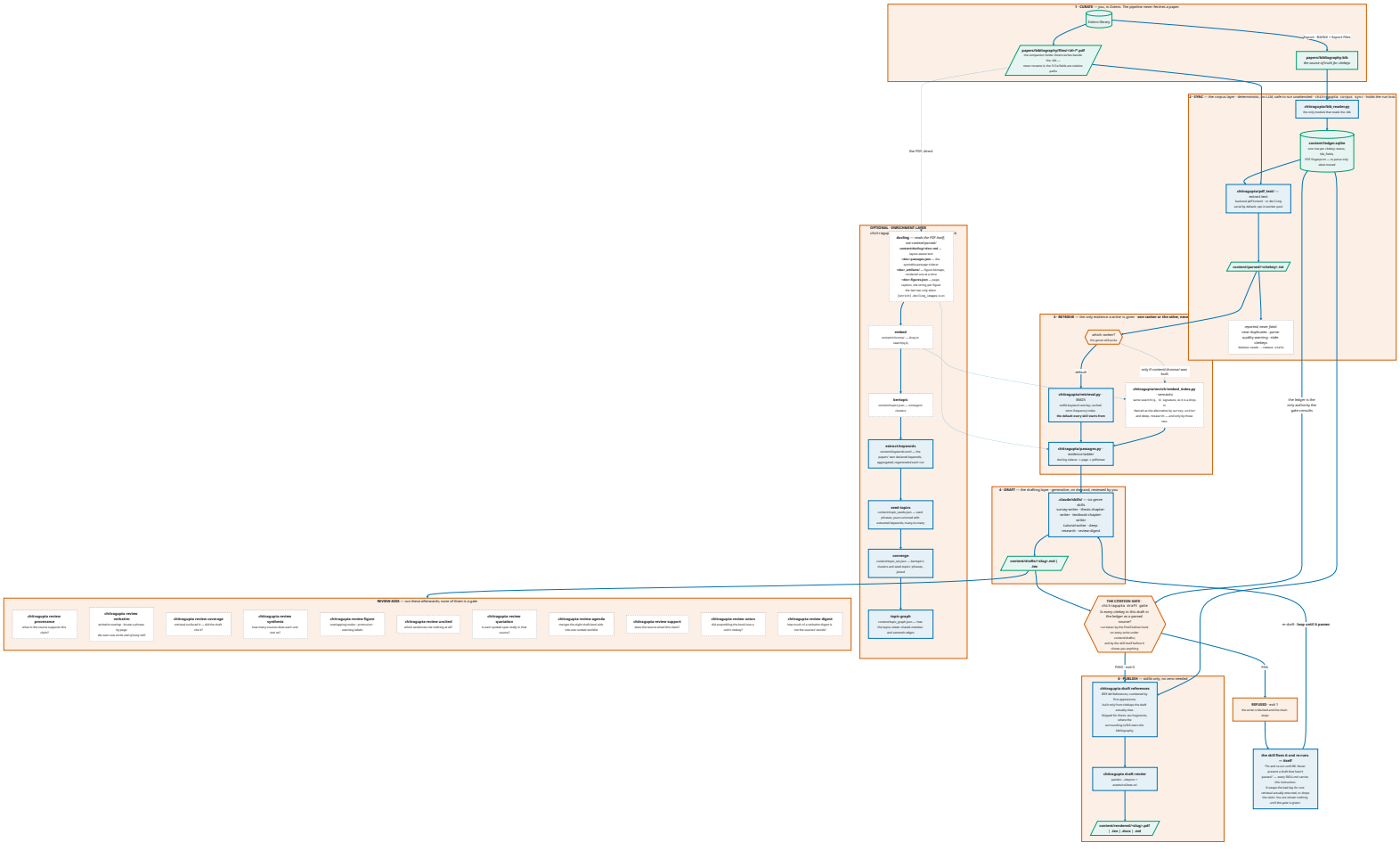
The full workflow — every module, every artifact


1 · CURATE — you, in Zotero. The pipeline never fetches a paper.
Zotero library
Format: BibTeX + Export Files
papers/bibliography/files/<id>/*.pdf
the acquisition folder — one numbered folder per item. the .bib … never rename it; the #id fields are relative paths
papers/bibliography.bib
the source of truth for citekeys
the PDFs direct
2 · SYNC — the corpus layer · deterministic, no LLM, safe to run unattended · v3.1 (+ pagelib · corpus · sync) · holds the run lock
chkraggraphs/bib_reader.py
the only module that reads the .bib
content/ledger.sqlite
one row per citekey: status, bib_fields, PDF fingerprint — re-parse only what moved
chkraggraphs/pdf_text/ — extract text
backend pdftotext · fitz · pdfplumber
quality gate: ≥ 200 words/page, ≥ 0.85 alpha
quarantine, never silent
content/parsed/<citekey>.txt
reported, never faked
never duplicate · garner
quality warning · stale
re-extract · force
skip
skip
skip