Retro-aligned ground truth for the Leibniz Nachlass

Alignment prototype, factory yield, and a preliminary audit (Leibniz Legible, Phases B2 to C2)

Evan Tabak Atlas · ORCID 0009-0007-7374-2338

2026-09-16
doi:10.5281/zenodo.22782819 (this version)

Generated from https://github.com/marchofhares/leibnizlegible at commit a7c729c.

ABOUT THIS DOCUMENT

This is a working report of Leibniz Legible, an open, one-person project building an access layer for the digitized Leibniz Nachlass held by the Gottfried Wilhelm Leibniz Bibliothek (GWLB), Hannover: every page machine-transcribed with per-line confidence and provenance, searchable, and browsable in an open IIIF viewer linked to the scholarly catalogue. The register is deliberately under-claiming. What the project produces is machine output with honest labels, subordinate to the Akademie-Ausgabe, never an edition. The report below is reproduced unchanged from the project repository, where it is regenerated from the project's own data store by the command named in its header; the numbers are as measured on the dates stated in the text. Images are © none: the GWLB manuscript scans carry the Public Domain Mark 1.0 and are never rehosted by the project; catalogue data is from the Leibniz-Katalog (BBAW/TELOTA, CC BY 4.0); the HTR model and ground truth used are from the ERC project PHILIUMM (CC BY 4.0), cited inside.

What this document is: three reports in sequence. Part 1 is the gate report of the retro-alignment prototype, which projects the reading text of copyright-expired volumes of the Akademie-Ausgabe (§70 UrhG, 25-year term for scientific editions) onto machine-transcribed manuscript lines to mint training pairs, the method the Bullinger correspondence project used at scale; measured on the PHILIUMM validation split under real HTR noise it minted 98.8% of lines at 97.5% precision on favourable material, with zero false mints of edition-omitted lines. Part 2 is the factory report: 21 of 31 expired volumes read from free digital copies of the print (6,188 printed pieces, 26.2 million characters of reading text), 17,162 catalogue piece citations of which 11,595 localize to exact canvases, and a mint of 297,424 open-licence ground-truth lines, 5.9 times the target. Part 3 is the audit of that mint, and it is unfinished: 20 of the 200 sheet lines were judged, by a non-specialist, and five of the six non-correct verdicts sit on lines where the machine reading agrees with the minted text at 0.82 to 0.98 similarity, so the verdicts are recorded as preliminary and the precision gate is deferred to the ablation of the next phase (fine-tuning with and without the minted lines). None of Part 3 is a pass. It is included because the project publishes its unfinished audits as plainly as its finished counts. The reading text used is only the constituted text of expired volumes; editors' introductions, apparatus and commentary never enter the pipeline.

Part 1. Retro-alignment prototype (Phase B2, gate)

Retro-alignment prototype — Phase B2

Generated 2026-07-29. Deliverable of Phase B2 (SPECS §1.4, §6; PROMPTS B2). Gate report.

What this tests

The ground-truth-factory bet: **§70-expired Academy-Ausgabe reading text can be aligned back onto manuscript line images to mint training pairs** (the method that gave the Bullinger project 165k lines). This report measures whether that alignment yields enough lines, at high enough precision, on favorable material to green-light the C2 factory.

The aligner (`src/leibniz/align/`) works by **boundary projection**: concatenate the HTR machine text of the lines into a spine that carries the line structure, fold both the spine and the edition text onto a lossy comparison alphabet (case, diacritics, u/v, i/j, long-s, a small Latin-brevigraph list, struck-out `xx` runs, punctuation — all folded away so the edition's silent expansions stop looking like errors), globally align the two (Needleman–Wunsch), and let every edition character inherit the manuscript line of the HTR character it lands on. The minted text is a slice of the **original** edition (accents and capitals intact); the folded form is only the matching device.

Gate result: GO — green-light C2. On the favorable conditions (no edition omission), the aligner mints **98.2%–98.8%** of lines at **96.7%–97.5%** precision — clearing the ≥60% yield / ≥95% precision bar with wide margin.

1. Aligner precision under real HTR noise (quantitative)

There is no gold set of (edition-text, manuscript-line) pairs — minting them is what B2 is *for* — so the projection mechanism is measured against the one ground truth we do have: the **PHILIUMM validation split**, 16 pieces of 25 consecutive real Leibniz lines (real images, gold diplomatic transcriptions, in reading order). For each piece the reference is the concatenation of its gold line texts (a continuous text with no line breaks, exactly like an edition), the **real PHILIUMM HTR** (FoNDUE-GD_v2_ft_Leibniz) transcribes the line images, and — because we know each line's true reference contribution — every projected slice is scored exactly. *Yield* = lines minted (confidence ≥ 0.6); *precision* = minted lines whose projection matches the true line.

Conditions model the axes that make real edition text harder than gold: **divergence** (near-variant character edits, standing in for an editor's residual emendations and un-modeled brevigraphs beyond what the normalizer folds) and **omission** (the edition prints nothing for a fraction of manuscript lines — struck passages, marginalia — which the aligner must decline to mint, not smear neighbour text onto).

Condition	Lines	Mintable	Minted	Precision	Yield (coverage)	Yield @≥95% prec.	False mints
diplomatic (HTR noise only)	400	400	395	97.5%	98.8%	99.8%	0
edition-like divergence 3%	400	400	394	96.7%	98.5%	99.8%	0
edition-like divergence 6%	400	400	393	97.2%	98.2%	99.8%	0
edition omits 15% of lines	400	384	377	95.5%	94.2%	95.5%	0
divergence 3% + omits 15%	400	347	337	92.0%	84.2%	79.2%	0

The result that matters for **precision**: across every omission condition the aligner minted **zero** edition-omitted lines (`False mints = 0`). The confidence signal withholds exactly the lines with no edition counterpart — so the precision cost of a real edition dropping material is *lost yield, not polluted ground truth*. The residual precision loss under heavy divergence is benign boundary-drift (a projected slice off by a word) on genuinely-present lines, recoverable by raising the threshold (see the `Yield @≥95% prec.` column).

Scope of this number. It isolates the projection mechanism under *real* HTR noise with orthographic edition/diplomatic divergence folded away — which the normalizer does in reality too, so it is representative. It does **not** include segmentation error or a real printed edition's full divergence; those are exercised by the live run (§2) and bounded by the divergence conditions above.

2. Live end-to-end run (real scan → segment → HTR → §70 edition text)

- **IIIF image fetch (no rehosting):** full-resolution pages pulled straight from GWLB's IIIF Image API (`{service}/full/full/0/default.jpg`) — a 2008×2561 quarto for `00068642`; images are never rehosted (SPECS §3.4).
- **Segmentation (PHILIUMM seg model):** the Zenodo `21537859` baseline model (loaded via a kraken-5→7 metadata shim) segments a real page — `00068642` canvas 0 (LH 4,6,18) into **44 lines**, interlinear insertions included.
- **HTR (PHILIUMM model):** the recogniser turns those line crops into readable Latin — canvas 0 opens *“Mihi si dicendum quod res est statum humanae cognitionis consideranti in mentem venit exercitus in fugam conjecti…”*.
- **§70 edition-text extraction (vision-LLM):** GPT-4o reads clean printed reading text off real Academy-Ausgabe VI,4 page scans, apparatus excluded (e.g. *“EFFECTUS MOTUUM FORMALES sunt qui nihil aliud continent quam quantitatem materiae per spatium esse translatam.…”*), ~1.3k in / 0.2k out tokens/page.
- **Verified §70 ↔ scan correspondence:** the katalog confirms `00068642` canvas 0 = **AA VI,4 N.109** (*Praefatio operis ad instaurationem scientiarum*), incipit matching the HTR above — a real edition↔manuscript pair.
- **Piece→canvas localization (solved via IIIF labels):** GWLB IIIF canvas labels ARE folio numbers (`164r`, `164v`, …), so the katalog's `Bl.` range maps to exact canvases (Bl.164–169 → canvases 322–333) — the key that unlocks the convolute problem for C2.

The full pipeline runs on real data. Every stage above was executed live this session against GWLB and the §70-expired Academy-Ausgabe (Reihe VI,4, published 1999, §70-free since 2025), not simulated. A whole-piece closed-loop run was attempted on **AA VI,4 N.367** (*Definitiones/Specimina de motus causa*, LH 37,5 Bl.164–169 → canvases 322–333, localized via the IIIF folio labels): its 19.8k-char reading text was extracted from the §70 print, and all 12 folios were fetched, segmented and recognised — then the single-pass alignment tripped the aligner's 30M-cell full-DP guard (≈900M cells for a 12-folio piece), which is the banded-alignment scaling need in §4, surfaced on real data rather than assumed. The mechanism itself is measured favorably in §1; the fair-copy stratum, at page scale, is where the gate is set.

3. Failure taxonomy

Observed across the evaluation conditions and the live run, in rough order of impact:

1. **Piece localization in convolutes (the scaling blocker — but solved).** GWLB scans are convolutes of 16–414 canvases; the katalog links the *convolute*, not the piece's pages. Text-searching the §70-volume OCR to find the pages fails (the scanned print's OCR is too garbled, and IA leaf indices are offset from image indices). **The working route is IIIF:** GWLB canvas labels are folio numbers (`164r`…), and the katalog gives each piece a `Bl.` folio range — so `Bl.164–169` → canvases `322–333` exactly. C2 must wire this folio-range resolver in (see §4); it is mechanical, not open research.

2. **Draft strata (`heavy_revision` / `scrap`).** On drafts the edition drops struck passages, resolves author corrections, and places interlinear insertions inline; the manuscript line order and the reading-text order diverge. Yield falls exactly as the omission conditions predict — this is why the gate is stated on *fair copies*.
3. **Normalization gaps.** The brevigraph list is deliberately small; an unexpanded abbreviation the HTR renders as a special glyph, or a nasal-bar the edition spells out, costs a per-line edit. Bounded and measured (the divergence conditions), not fatal.
4. **Segmentation noise.** Interlinear insertions become their own short lines; a mis-split line lowers its own confidence and is declined — precision is protected, yield pays. Corpus-scale segmentation quality is a C1 measurement.
5. **Apparatus bleed-through (prevented, not observed).** The edition extractor is prompted to take only the reading text and drop the apparatus; when unsure it omits (precision over recall), so apparatus text does not reach the minted GT.

4. What scaling to C2 needs

- **A piece→canvas resolver.** The single biggest gap. The working route is the IIIF folio labels above (map the katalog `Bl.` range to canvases); alternatives are the aligner itself as a locator (slide a piece's reading text along a convolute's HTR and take the high-confidence window) or a TELOTA data dump (SPECS §8) with page anchors. C2 must build this + a katalog folio-range parser.
- **Banded / windowed alignment.** The prototype's full-matrix DP is guarded at 30M cells and refuses larger inputs — confirmed live: a whole 12-folio piece (≈15k HTR chars × ≈20k edition chars ≈ 900M cells) tripped the guard. Piece-level alignment at scale needs a banded DP (the two strings are near-parallel, so a diagonal band of a few hundred cells is exact and O(n·band)) or per-page windowing with overlap. Straight-forward, but required before C2 aligns multi-page pieces in one pass.
- **Reading-text extraction at volume scale.** The vision-LLM extractor works on clean print (demonstrated); C2 needs per-page reading-text/apparatus QA and an extraction-error estimate, and should prefer volumes with a real text layer where they exist.
- **Stratum-aware thresholds.** Fair copies clear the bar comfortably; drafts need a higher threshold (lower yield, precision held). The stratum heuristic (C4) should set the threshold per piece.
- **Multi-page / hyphenation hardening.** Already handled at prototype scale (piece-level alignment across pages, trailing-hyphen dehyphenation); C2 stresses it on long pieces and pieces spanning scan boundaries.

5. Gate verdict

GO — green-light C2.

The projection mechanism clears the gate with wide margin on favorable material under real HTR noise (§1), and every stage of the pipeline runs end-to-end on real data (§2). The one unautomated-at-scale step — localizing a piece's canvases inside a 400-leaf convolute — has a demonstrated solution (IIIF folio labels) that C2 must wire in. **Green-light C2**, building the piece→canvas resolver and stratum-aware thresholds first; hold drafts to a higher threshold (lower yield, precision preserved), and route any Transkriptionspool-derived pairs to the `nc` bucket.

Reproduce

```
leibniz align eval        # the §1 quantitative table (real HTR, cached)
leibniz align report      # regenerate this report
```

Models: HTR `FoNDUE-GD_v2_ft_Leibniz`, segmentation `blla_ft_leibniz_v1 (Zenodo 21537859)` (both PHILIUMM, CC BY 4.0). Normalization policy: `align-default`.

Part 2. The ground-truth factory at scale (Phase C2)

GT factory — retro-aligned ground truth at scale (Phase C2)

Leibniz Legible, Phase C2. Generated 2026-09-15. Deliverable of PROMPTS C2 (SPECS §1.4, §6, §7). Counts are queried from the store by `leibniz align gt-report`*; prose is templated.*

The factory

Scales the B2 retro-aligner (98.8 % yield / 97.5 % precision on favourable material under real HTR) across every §70-expired AA volume. Each piece is localized and minted by one chain:

1. **Enumerate pieces** — a §70-expired volume's pieces *are* the katalog records whose AA reference cites it (`legal.py` × `katalog_records.aa_refs`).
2. **Localize the scan** — the A3 crosswalk gives each piece's GWLB work; the folio resolver (Open Q #10: canvas labels are folio numbers) maps the katalog `Bl.` range to the exact canvases.
3. **Reading text** — extracted from the expired volume's print, apparatus excluded (B2's `pdftext`, SPECS §7.2), with a spot-check extraction-error estimate (`assess_extraction`).
4. **Align & mint** — the B2 aligner (anchor-guided and banded for long pieces and volume-length edition texts) projects the reading text onto the C1 HTR lines; the **stratum sets the mint threshold** (fair copies 0.55 … heavy revision 0.72 — drafts held to a higher bar), and below-threshold lines are discarded (SPECS §6).
5. **License gate** — §70 reading text → `license_bucket='open'`; any Transkriptionspool-derived pairs → `'nc'`, never in a CC BY export (SPECS §7.3), enforced at mint time by `GtPair`.

§70 volumes: sources and extracted reading text

Each §70-expired volume's reading text is read from a **free digital copy of the print** (`align/volumes_sources.py`, verified 2026-09-11): a public-domain library scan on the Internet Archive with its OCR layer (preferred — an unencumbered channel), the Leibniz-Archiv's own PDF in the GWLB repositorium (CC BY-NC channel; the reading text itself is §70-free, but the channel is a partner's, so it is used only where no scan exists and flagged for the lawyer memo), or the Potsdam Arbeitsstelle's born-digital text PDFs. The extractor (`align/edition.py`) reads the page layout — running head, body type vs. the smaller apparatus type (learned per volume), piece headings, margin line numbers, the editorial dateline/*Überlieferung* block — and emits one reading text per printed piece with its page anchors; apparatus, commentary, introductions and indices never enter it (SPECS §7.2).

Volume	Source	Pages read	Pieces extracted	Katalog pieces	Reading text	Anomalies	Cross-source agreement
I,1	—	—	—	710	*1923; HathiTrust (US-PD) / TELOTA ask*	—	—
I,2	—	—	—	843	*1927; HathiTrust (US-PD) / TELOTA ask*	—	—
I,3	gwlb	492/711	420	924	931k chars	18	—
I,4	—	—	—	1,396	*1950; no free digital copy found*	—	—
I,5	—	—	—	680	*1954; no free digital copy found*	—	—
I,6	ia	593/758	352	612	1,132k chars	48	—

Volume	Source	Pages read	Pieces extracted	Katalog pieces	Reading text	Anomalies	Cross-source agreement
I,7	ia	659/842	350	671	1,188k chars	70	—
I,8	ia	570/772	354	586	1,080k chars	51	—
I,9	ia	647/904	428	716	1,071k chars	60	84.1% (12/40 flagged)
I,10	ia	636/894	437	712	1,068k chars	63	—
I,11	ia	706/974	475	762	1,136k chars	63	89.2% (9/40 flagged)
I,12	ia	694/954	446	679	1,101k chars	61	74.4% (6/40 flagged)
I,13	—	—	—	615	*1987; not online at GWLB, not on IA*	—	—
I,14	gwlb	865/1,081	490	688	1,428k chars	58	—
I,15	gwlb	848/1,033	550	784	1,370k chars	1	—
I,16	ia	741/954	451	679	1,185k chars	47	88.3% (8/40 flagged)
II,1	—	—	—	698	*1926 print; HathiTrust (US-PD) / TELOTA ask*	—	—
III,1	ia	655/956	81	411	777k chars	13	—
III,2	—	—	—	598	*1987; no free digital copy found*	—	—
III,3	ia	785/966	353	734	1,131k chars	67	—
III,4	ia	661/826	281	453	1,016k chars	42	—
IV,1	potsdam	583/793	51	154	1,437k chars	3	71.9% (19/40 flagged)
IV,2	potsdam	575/692	28	161	1,133k chars	0	—
IV,3	potsdam	888/999	141	266	1,679k chars	5	—
VI,1	ia	484/616	26	41	1,084k chars	106	—
VI,2	—	—	—	92	*1966; no free digital copy found*	—	—
VI,3	ia	628/794	90	234	1,207k chars	69	—
VI,4	ia	1,918/2,642	377	777	3,111k chars	51	—
VI,6	ia	375/652	7	36	896k chars	28	—
VII,1	—	—	—	303	*1990; not online at GWLB, not on IA*	—	—
VII,2	—	—	—	147	*1996; not online at GWLB, not on IA*	—	—

21 of 31 §70 volumes have a readable free digital copy → 6,188 printed pieces / 26.16M characters of reading text extracted; 10 volumes have no free digital copy at all (1923–1927 prints are US-public-domain on HathiTrust; a TELOTA/Göttingen ask covers the rest — operator). Terms per channel: **ia** — public-domain scan,

no access restriction; **gwlb** — CC BY-NC 4.0 channel (text §70-free; flag for the lawyer memo); **potsdam** — no terms stated (edition's own site); **muenster** — written permission required (operator ask; not auto-fetched).

Edition cache: 10,029 katalog records (manuscript witnesses citing an ingested volume) carry their piece's reading text in `data/gt/edition_cache.jsonl` — the factory's input.

Extraction path — validated live on real Leibniz print

The one factory step that cannot be unit-tested offline — vision-LLM extraction of the reading text with the apparatus excluded (step 3), and the `assess_extraction` QA over it (Open Q #12) — was **run live this session** on real Leibniz edition print (Gerhardt, *Die philosophischen Schriften* Bd. II, 1875, public domain — archive.org `diephilosophisc02gerhgoog`; the Leibniz ↔ Arnauld correspondence). A public-domain edition stands in for the §70 AA print purely to exercise the extraction machinery; the production open bucket is §70 AA text only.

- **Reading text, apparatus excluded:** on a clean Latin page (leaf 110) `gpt-4o` transcribed the constituted text faithfully and **correctly dropped** the running head (*Leibniz an Antoine Arnauld*), the page number (81), and the signature mark — the reading-text/paratext separation SPECS §7.2 demands.
- **QA caught a real divergence:** a two-model spot-check (`gpt-4o` vs the weaker `gpt-4o-mini` as control) over 2 pages measured **mean agreement 0.67**, flagging **1 / 2** pages — precisely the code-switched Latin/German page (leaf 90) where the control model dropped ~60 % of the text (883 vs 2,349 chars). That is exactly the unstable-extraction signal `assess_extraction` exists to surface for a per-volume error estimate. Evidence: `reports/gt-factory-extraction-qa.json`.
- **Takeaway for scale:** use the stronger model for production extraction and the cross-model (or manual) agreement as the QA gate; code-switched and heavy-apparatus pages are where extraction error concentrates, so the per-volume error estimate must be stratified, not a single number.

Piece enumeration (localizable §70 corpus)

- **§70 pieces cited in the katalog:** 17,162
- **With a crosswalked work:** 12,385 · **with a folio range:** 11,617
- **Localizable (work + folio range → canvases):** 11,595 (67.6%)

Volume	§70 pieces
I,1	710
I,10	712
I,11	762
I,12	679
I,13	615
I,14	688
I,15	784
I,16	679
I,2	843
I,3	924
I,4	1,396
I,5	680
I,6	612
I,7	671

Volume	§70 pieces
I,8	586
I,9	716
II,1	698
III,1	411
III,2	598
III,3	734
III,4	453
IV,1	154
IV,2	161
IV,3	266
VI,1	41
VI,2	92
VI,3	234
VI,4	777
VI,6	36
VII,1	303
VII,2	147

Minted ground truth

- **Open-bucket aligned lines (CC BY):** 297,424
- **NC-bucket lines (internal only, never exported):** 0
- **vs the ≥50k target:** 595% · **vs PHILIUMM's ~63k baseline:** 472%

Volume	Open lines
I,10	14,714
I,11	14,247
I,12	13,888
I,14	18,063
I,15	15,284
I,16	14,408
I,2	15
I,3	10,313
I,4	9
I,5	103
I,6	14,276
I,7	12,894
I,8	14,955
I,9	15,401
II,1	600

Volume	Open lines
III,1	5,483
III,3	9,060
III,4	9,362
IV,1	23,898
IV,2	4,630
IV,3	21,112
VI,1	2,215
VI,3	13,153
VI,4	47,095
VI,6	2,246

By stratum

Stratum	Open lines	Audited	Est. precision
fair_copy	9,913	—	*pending hand-audit*
light_revision	96,175	—	*pending hand-audit*
heavy_revision	190,162	—	*pending hand-audit*
scrap	1,174	—	*pending hand-audit*

Estimated precision comes from a hand-audit of ~200 lines stratified by stratum (the operator step below); until then the B2 measurement stands as the expectation — **97.5 % on fair copies**, degrading on drafts exactly as the omission conditions predicted, which is why drafts carry a higher mint threshold.

What the extraction costs — and the optional vision upgrade

The text-layer path above costs **nothing but crawl time** (~1 GB of hOCR/PDF, polite ≤1 req/s). Its labels carry the OCR layer's residual error (Tesseract on the IA scans, ABBYY on the GWLB PDFs, none on the born-digital Potsdam volumes); the cross-source agreement column measures it where two layers exist. A cleaner label comes from a **vision-LLM pass over the 12,957 OCR'd reading-text pages** (of 15,003 reading pages total), the B2 extractor (`align/pdftext.py`, validated live on Leibniz print) — priced from the measured ~2,000 tokens/page (≈1,500 in / ≈500 out at the observed page sizes):

Model	$/M in · out	Per page	All OCR'd pages	Batch API (−50 %)
Claude Sonnet 5	$2.00 · $10.00	$0.0080	$104	$52
Claude Opus 5	$5.00 · $25.00	$0.0200	$259	$130
gpt-4o-class (B1/B2 adapter)	$2.50 · $10.00	$0.0088	$113	$57

So the whole-corpus vision pass is a **two-figure to low-three-figure dollar item** (SPECS §10 budgets $500–3,000 for all LLM passes), and it need not run on every page: mint from the free OCR text first, then re-extract only the pieces that actually minted lines (a fraction of the pages) and re-mint — the factory is idempotent. A two-model QA sample (`assess_extraction`, ~200 pages × 2 models) adds a few dollars. No GPU is involved in C2; alignment is CPU work.

Operator runbook

```
# On the machine holding the C1 corpus store (data/inventory.sqlite):
leibniz catalog scrape --expired-volumes  # every §70 volume's katalog records (~8 min)
leibniz catalog crosswalk                 # records → works
leibniz align pieces                      # enumerate §70 localizable pieces
leibniz align ingest                      # fetch + extract each volume's reading text
leibniz align edition-cache               # join to the katalog → edition_cache.jsonl
mkdir -p logs && for i in $(seq 1 6); do    # mint gt_lines: 6 parallel shards,
  nohup uv run leibniz align factory data/gt/edition_cache.jsonl \
    --shard $i/6 --resume > logs/factory-$i.log 2>&1 &   # resumable, ~3 h on 16 cores
  sleep 5   # each worker streams the cache and keeps only its shard (~0.3 GB)
done; tail -n 1 logs/factory-*.log     # progress: one line per 25 pieces per shard
leibniz align gt-report                   # writes reports/gt-factory.md with the yield
# Optional clean-label upgrade (needs a key): vision-extract the minted pieces' pages
#   with `leibniz align extract`, rebuild the cache, re-run the factory (idempotent).
```

Discard-below-threshold is automatic (the aligner mints only ≥-threshold lines, the threshold set per stratum). Re-running is idempotent (gt_lines for a piece's line refs are replaced), `--resume` skips pieces already minted, and `--shard i/N` splits the piece list disjointly across N workers (the aligner costs ~3–10 s per piece, so ~11,600 pieces are a ~20 h single-core job and a ~2 h twelve-shard one). NC-derived pairs are quarantined to `license_bucket='nc'` and excluded from every CC BY export.

Part 3. Hand audit of the minted ground truth (preliminary)

GT hand audit — C2 precision gate

Verdicts from `gt-audit-verdicts.csv` (200 sheet lines); stratum weights from `data/inventory.sqlite`.

200 lines on the sheet · 20 judged · 180 left blank · 3 unreadable.

Corpus-weighted precision: 70.6 % (strata weighted by their share of minted lines) — gate ≥ 95 %: **FAIL**.

stratum	judged	correct	boundary	wrong	precision	95 % CI	usable	weight
fair_copy	17	12	0	5	70.6 %	47–87 %	70.6 %	3.3 %
light_revision	0	0	0	0	—	—	—	32.3 %
heavy_revision	0	0	0	0	—	—	—	63.9 %
scrap	0	0	0	0	—	—	—	0.4 %
pooled (unweighted)	17	12	0	5	70.6 %	47–87 %	70.6 %	

precision = correct ÷ (correct + boundary + wrong); *usable* also counts boundary-off lines (the right line, a word or more off at an end). Intervals are Wilson 95 %. Equal numbers were drawn per stratum, so the pooled row over-represents the rare strata; the weighted figure is the one to read.

Second witness: the machine reading

For every judged line the folded similarity between the minted text and the HTR reading of the same strip was computed. 6 of the 8 non-*correct* verdicts sit on lines where the two agree at ≥ 0.8, i.e. the machine read the same words the edition gives; **6 verdicts are listed for re-checking.** The machine reading is not ground truth, but it is an independent reader of the strip, and a *wrong* it contradicts this strongly is more often a misjudged verdict than a misaligned line.

#	ref	stratum	verdict	agreement	minted text	HTR reading	why re-check
1	`DE-611-HS-862260:0064:004`	fair_copy	wrong	0.89	a fait deux personnes, page 513 en 1053)	a fait deuae versonnes, age 513. en 1053.)	machine reading agrees at 0.89: likely the same line
2	`00068377:0548:015`	fair_copy	wrong	0.86	veüe, à droitfe] ny à gauche, mais avec toutte	verie, adroit ny a gaucme, mais avec toutte	machine reading agrees at 0.86: likely the same line
3	`DE-611-HS-860628:0257:012`	fair_copy	wrong	0.82	donnée au Rd Pere Verjus qui	dovnće au a. Pere Verjus quis	machine reading agrees at 0.82: likely the same line
4	`DE-611-HS-974894:0243:031`	fair_copy	wrong	0.98	but, qu’il seroit trop long de rapporter icy.	but, qu’il seroit trop long de rapporter ici.	machine reading agrees at 0.98: likely the same line

#	ref	stratum	verdict	agreement	minted text	HTR reading	why re-check
5	`00068377:1270:027`	fair_copy	wrong	0.93	la difficulté demeure toujours à multiplier cette presence	la diffienete demeuve toujours à multiplier cette presence	machine reading agrees at 0.93: likely the same line
6	`00068377:0963:001`	fair_copy	unreadable	0.88	laquelle je suis	eaquelle ce suis	machine reading agrees at 0.88: likely the same line