Leibniz Legible: an open access layer for the digitized Leibniz Nachlass

Project statement, September 2026 (version 1: the corpus machine-read; search and viewer in public beta)

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

2026-09-16
doi:10.5281/zenodo.22782813 (this version)

Generated from https://github.com/marchofhares/leibnizlegible at commit ede16e1.

Abstract

The Leibniz Nachlass in Hannover, some 236,000 digitized page images, is the largest body of writing by any early-modern philosopher that remains mostly unread: after a century of work the Akademie-Ausgabe has printed roughly a quarter of it. Leibniz Legible is a one-person open project that builds an *access layer* over those images rather than an edition: every page machine-transcribed with a per-line confidence score and a full provenance record, searchable with tolerance for early-modern orthography, and browsable in an open IIIF viewer that loads the library's own images and shows the machine reading beside them with honest labels. This statement records the state of the project in September 2026. The digitized corpus has been counted for the first time from the library's own metadata (2,225 works, 236,795 page images). The character error rate of the ERC PHILIUMM project's Leibniz handwriting model has been reproduced under a frozen protocol (7.95%, 95% CI 7.49–8.46, against a claimed 8.33%), and the same protocol gives the first published numbers for frontier vision-language models on Leibniz's hand (37% to 79% character error rate, zero-shot). The whole corpus has been segmented and read with that model: 236,210 pages, 13.5 million lines, each with a confidence and a run record, on one desktop computer in six weeks. A ground-truth factory has aligned the reading text of copyright-expired edition volumes back onto the machine lines and minted 297,424 training pairs, six times the target, at a precision that is measured on favourable material but only preliminarily audited on the corpus. The serving layer, a search index, a JSON API, IIIF Presentation 3 manifests with W3C annotations, and a viewer, is built, and the datasets are packaged with cards that carry their provenance, licence and error rates. The statement closes with what the next model can and cannot buy, what the crowd and language models can and cannot fix, and the roadmap to the public release.

1. What this is, and what it is not

Gottfried Wilhelm Leibniz left about 50,000 pieces on about 100,000 sheets. The Gottfried Wilhelm Leibniz Bibliothek (GWLB) in Hannover digitized the whole Nachlass between 2016 and 2019 and serves it openly, page by page, under the Public Domain Mark. The Akademie-Ausgabe, the critical edition begun in 1923, has printed a part of it with exemplary care, and will continue to do so for decades. The BBAW's Ritter-Katalog, more than 70,000 records, describes almost every piece but transcribes very few. So a scholar who wants to know whether Leibniz wrote a given phrase, or where a given correspondent appears, can look at any page in the world but cannot search a word of most of them.

Leibniz Legible closes that gap in the only way a small project honestly can: with machine reading, labelled as machine reading. Its posture is fixed in its specification and repeated on every surface it ships. It is an access layer of *Vorausedition* grade, explicitly subordinate to the Akademie-Ausgabe; it is machine output with honest labels,

never "an edition"; the register toward the institutions that hold and edit the material is that of a supplier and admirer, never a competitor; and when in doubt, the project under-claims.

The deliverables are correspondingly plain: an inventory of every page image with its catalogue record; a full-corpus machine transcription with per-line confidence and provenance; a retro-aligned ground-truth dataset; a fine-tuned recognition model; a reproducible benchmark harness; one modest web application with a search box and a viewer; IIIF manifests and annotations so that other viewers can consume the layer; and reports. Code is Apache-2.0; data and reports are CC BY 4.0; the inventory is CC0. Images are never rehosted.

2. The state of the work, September 2026

The project runs as a sequence of gated phases (Table 1). Everything below is measured, and every number is regenerated by a command from the project's own data store; the three companion reports deposited alongside this statement carry the detail.

Table 1. Phases and gates.

Phase	What	Gate	State
A1	OAI-PMH/IIIF harvest, corpus census	first real page count	done: 2,225 works, 236,795 pages
A2	image cache		done: 236,779 pages, 395.6 GB
A3	catalogue crosswalk		done for the §70 volumes; full scrape is an operator job
B1	benchmark harness, PHILIUMM reproduction	CER within 1 point of the claim	passed: 7.95% vs 8.33%
B2	retro-alignment prototype	yield ≥60% at precision ≥95%	passed: 98.8% at 97.5%
C1	corpus segmentation and recognition		done: 99.75% of pages, 13.5M lines
C2	ground-truth factory	≥50k open lines	done: 297,424 lines; audit preliminary
C3	fine-tune v2	CER ≤7% on Latin/French	next
C4	corpus re-run, language id, strata		after C3
D1–D3	search, viewer, releases		built on v1; public beta

2.1 The corpus, counted

The census (Phase A1) is, to our knowledge, the first published page-image count of the Hannover Nachlass computed from the library's own METS metadata and cross-checked against IIIF manifests: 2,225 works and 236,795 page images across five collections. The annotated printed books of the Marginalien are 44% of the pages; the correspondence 31%; the manuscripts proper 25%. A finding that revised the plan: only 812 works (77,633 pages, 33%) are served through a IIIF Image API; the rest expose static JPEG derivatives only, so the viewer degrades to plain images where no tile service exists, and the image cache pulled the uniform METS delivery derivative for every page instead.

2.2 The model, reproduced before anything was built on it

In July 2026 the ERC project PHILIUMM (Paris) released a Kraken recognition model fine-tuned for Leibniz's hand, a segmentation model, and about 63,000 lines of ground truth, all CC BY 4.0. Phase B1 built an engine-agnostic evaluation harness with a frozen protocol (Levenshtein on Unicode code points, a named normalization policy that mirrors the model's own training normalization, micro-averaging, bootstrap confidence intervals) and

measured the model on its own 1,878-line validation split: character error rate 7.95% (95% CI 7.49–8.46) against the claimed 8.33%, word error rate 27.0%, one line in four character-perfect. The claim reproduced, and the project built on the model.

The same harness scored frontier vision-language models zero-shot on a 150-line subsample. They are not close: 36.6% (GPT-4.1) to 78.7% (the largest reasoning models) character error rate against 8.2% for the fine-tuned model on the same lines, the larger models most often "modernizing" the archaic spelling. A purpose-built open model reaches about 29% on handwritten Latin in a recent benchmark. These are, as far as we know, the first published numbers for language models on seventeenth-century secretary hands, and they settle a question the field keeps asking: for this material, a small fine-tuned recognizer beats a general model by a factor of four to ten.

2.3 The whole Nachlass, read once

Phase C1 built a resumable, idempotent batch pipeline (segment, then recognize, with a status machine over pages and one run record per batch carrying model, parameters, counts and git commit) and ran it over the entire corpus: 236,210 of 236,795 pages recognized, 13,508,625 lines with text, confidence and provenance. The remainder is enumerated, not lost: 569 pages skipped with a recorded reason (blank, specks, oversize foldouts, degenerate geometry) and 16 pages behind broken delivery URLs at the library. The run took about six weeks of wall-clock time on one sixteen-core desktop with a single consumer GPU, in the end as four sharded CPU segmentation workers alongside one GPU recognizer. Segmentation, the known risk on layered drafts and marginalia, was measured rather than assumed: every page carries a statistics row (line count, region coverage, line-height variation, overlapping boxes, short lines), and 92% of pages have overlapping line boxes, which is the layout signature of revision.

Per-line confidence is real CTC posterior mass, checked at scale (mean 0.85, median 0.91; 53% of lines at or above 0.9, 6% below 0.5). It is the signal the viewer shades and the signal a review queue would draw on.

2.4 Ground truth from the edition, at scale

The strategy for improving the model is the one the Bullinger correspondence project demonstrated: align the reading text of a printed edition onto machine-transcribed lines and keep the pairs that match. German copyright gives scientific editions a 25-year term (§70 UrhG), so the reading text of every Akademie-Ausgabe volume published up to 2000 is free today, and a new volume or two becomes free each January; editors' introductions, apparatus and commentary are ordinary works and never enter the pipeline.

Phase B2 built the aligner: both texts are folded onto a lossy comparison alphabet (case, diacritics, u/v, i/j, long-s, ligatures, a small list of Latin brevigraphs, punctuation), aligned globally, and every edition character inherits the manuscript line of the machine character it lands on, so the minted text is a slice of the original edition, accents and capitals intact. Measured under real machine noise on the PHILIUMM validation split, the aligner mints 98.8% of lines at 97.5% precision on clean material and, in every condition where the edition omits lines, mints zero edition-omitted lines: the confidence signal withholds exactly the lines with no counterpart, so precision loss is lost yield, not polluted ground truth.

Phase C2 scaled it. The catalogue's edition references give every piece printed in an expired volume; the catalogue's folio ranges map to exact canvases because the library's IIIF canvas labels are folio numbers; the reading text of 21 of the 31 expired volumes was read from free digital copies of the print (public-domain library scans with their OCR layer, the library's own PDFs, born-digital PDFs) by a layout-aware extractor that learns each volume's body and apparatus type sizes and peels running heads, margin numbers and datelines; and an anchor-guided banded aligner handles pieces whose edition text runs to three quarters of a million characters in a few megabytes of memory. The mint produced 297,424 open-licence ground-truth lines from 6,383 catalogued

pieces in about two hours on six workers: 9,913 from fair copies, 96,175 from lightly revised pages, 190,162 from heavily revised drafts, 1,174 from scraps, each with its alignment confidence and its stratum, drafts held to a higher threshold than fair copies.

What the factory cannot measure is precision on the corpus, and here the statement must be exact. A 200-line hand audit was tooled and begun; the operator judged 20 lines, all fair copies, and stopped, saying so plainly. The verdicts were 12 correct, 5 wrong, 3 unreadable, which as written fails the 95% gate; but every one of the five "wrong" verdicts sits on a line where the machine's own independent reading agrees with the minted text at 0.82 to 0.98 similarity, which reads as misjudged verdicts rather than misaligned lines. None of that is a pass. The audit stays open for a reader of the hands, and the next phase's ablation (training with and without the minted lines) is the empirical test.

2.5 The serving layer

Phases D1 and D2 are built and enter public beta with the v1 transcription. The search index folds both the indexed text and the query onto the aligner's comparison alphabet, so a search for *ut* finds *vt* and a search for *Jesus* finds *Jeſus*, and it runs on either SQLite FTS5 (one file, prefix matching) or Meilisearch (typo tolerance). The API returns pages with their lines, geometry, text, confidence, status and provenance, and the viewer shows an OpenSeadragon canvas loading tiles directly from the library's Image API (or the static derivative), a polygon overlay of the lines, and a line panel with a status badge, a confidence band and a provenance disclosure for every line, in English and German. Each work is also served as a IIIF Presentation 3 manifest whose canvases reference W3C annotation pages carrying the transcription as *supplementing* annotations, with the provenance tuple embedded under a project namespace, so Mirador and other viewers see the same labels. Every surface carries the same three attribution lines and the same honesty banner.

Phase D3 packages the datasets (inventory, transcriptions, ground truth) as Parquet with a manifest of checksums and a card per dataset that states schema, provenance chain, licence, attribution, measured error rates and the anti-contamination note: machine output, do not ingest as verified text. Uploads follow a checklist; the reports, including this statement, are the first deposits.

3. The honesty regime

Four rules run through the code and are worth stating as design, because they are what makes machine reading of a philosopher's papers defensible.

Every line carries its provenance. The store, the API, the annotations and the exports all carry, for every line, the image URI and region, the model and version, the run and its date, the confidence, and a status from a four-word vocabulary: *machine* (unreviewed recognizer output), *aligned* (matched to a printed edition), *corrected* (a human or crowd correction, kept as a new row, the machine row immutable), *verified* (checked by a named expert). Nothing is published stripped of these fields.

Licence buckets are enforced at write time. Only the constituted reading text of copyright-expired volumes is ever redistributed; text derived from non-commercially licensed sources (the library's transcription pool, its repository PDFs) may be used internally for alignment and evaluation under the text-and-data-mining provisions but lands in a quarantined bucket that no export and no public page can reach.

Strata are first-class. Fair copies, light revision, heavy revision and scraps are labelled and reported separately, because an aggregate error rate hides the failure mode that matters. The same holds for languages: the 7.95% is a Latin-and-French number by construction; German and Kurrent, about 15% of the corpus, are transcribed by the same model and are expected to be far worse, and they will be measured and reported as such, whatever the number.

The register under-claims. The word "edition" is never used for the project's output; "agreement" is used where a crowd converges, never "accuracy"; the viewer's banner says errors are expected and gives the rate.

4. What the numbers mean

A character error rate near 8% is neither a triumph nor a failure; it is a particular kind of usefulness, and the literature is clear about which. For scholarly reading, the convention treats below 5% as very good, 5% to 10% as good, and above 10% as poor, the point past which correcting the machine costs more than transcribing afresh. For search, exact-match retrieval degrades measurably from about 5% error, while keyword spotting survives to about 30%, and it is the *word* error rate, here 27%, that predicts what a searcher experiences; the index's orthographic folding and typo tolerance exist to recover much of that loss. For downstream language processing, word accuracy above 90% is the usual comfort line and 80% the floor. For retrieval-augmented question answering over noisy text, the tolerance is lower still: recent benchmarks put French at "high noise" above about 5.6% character error, and generation quality falls with it.

So the v1 layer buys a good reading surface and a workable search index, and it does not buy reliable machine question-answering over the Nachlass. That is why the viewer always shows the image beside the text, why every citation the API gives is a line identifier and an image region, and why any future reading layer that normalizes or corrects the text will be a second, separately labelled layer on top of the diplomatic one, never a replacement for it.

5. What the next model can buy, and what it cannot

The next phase fine-tunes a second model on PHILIUMM's ground truth plus the 297,424 minted lines. It is worth being candid about the expected size of the gain, because the closest published analogue runs against the optimistic case. The Bullinger project went from about 110,000 alignment-minted lines to about 377,000 and its character error rate did not move (9.13% to 9.1–9.2%); its own refined ground truth measures 6.5% error against manual annotation, and that label floor, not model capacity, is what binds. Our labels are edition reading text, which silently expands abbreviations, normalizes *u/v* and *i/j* and omits struck passages, scored against a diplomatic validation set; each expansion is a systematic insertion error, and systematic error does not average out with scale. The mint also keeps only lines the first model already read well, so the new data is biased toward what the model already knows. The honest expectation is about one point, from just under 8% to about 7%, with four cheap changes that move the odds: re-extract the edition pages with a vision model to remove OCR-layer noise before training; stage the training so the hand-corrected diplomatic lines come last and weigh most; filter the minted lines by the recognizer's own confidence as well as by alignment similarity, and report the composition of the remaining errors; and decode with a character n-gram language model trained on diplomatic text, which in a comparable system removes about a tenth of the errors for free.

Two larger points follow. First, the corpus-wide gain will likely exceed what the gate shows, because most minted lines come from the correspondence, including letters in correspondents' hands the validation set never sees; a corpus-representative held-out set will report that separately. Second, no model reaches zero. Professional transcribers of these hands are measured at 1% to 13% error depending on skill and material, and the best ground truth in the genre carries a few percent of its own. The target is therefore a diplomatic layer near 6–7% with calibrated confidence, plus targeted human and machine review where it matters: confidence-ranked queues, disagreement between models as the flag (a model cannot see its own errors; two different readers can), and image-grounded correction only as a second labelled layer. On the validation split, correcting the worst tenth of lines would halve the residual error; the worst fifth, cut it to a third. Concentration, not volume, is the lever.

6. Roadmap

1. **C3, fine-tune v2** with the four changes above, a page-disjoint diplomatic held-out set, and the ablation that doubles as the test of the minted ground truth. Ship the best model whatever the number, with a model card that reports per-stratum and per-language rates.
2. **C4, corpus re-run and enrichment**: every page re-read by v2 under a new run (the old rows kept as provenance), per-line language identification, a stratum heuristic on every page, and a corpus report whose numbers (the real language split of the Nachlass, the real stratum distribution) nobody has computed.
3. **Public beta now, on v1.** Nothing in search or the viewer depends on the second model; the swap is a re-index. The serving layer is where the project becomes real to the library, the editors and any funder, and where usage begins to say which lines matter.
4. **Releases.** The inventory (CC0), the transcriptions and the ground truth (CC BY 4.0) as Zenodo datasets with cards, mirrored on Hugging Face; the model as a Zenodo model record after C3; code releases archived from GitHub. The census, the reproduction and the retro-alignment reports are deposited with this statement.
5. **Relationships.** The project seeks cooperation with the GWLB (whose own recognition pilot targets 2027 and whose master scans are offered to scientific projects on request), with PHILIUMM (whose models just read the whole Nachlass), with TELOTA at the BBAW (a catalogue dump would replace polite scraping), and with the Leibniz-Archiv and the Leibniz-Gesellschaft, in the subordinate-Vorausedition framing throughout. Where the law permits but a partner would wince, the project asks first.

7. Companion projects

Calculemus is a citizen-science game: a side-scrolling runner in which the player adjudicates A/B transcription candidates over fragments of the Nachlass, with calibrated player reliability, gold items with known answers, and a per-fragment posterior that retires a line when the crowd converges. Its output is crowd *agreement* over machine text, exported into Leibniz Legible as *corrected* lines, never as *verified* ones, and it is designed to run against the very lines a search hit or a reference query depends on. Its statistics and its playable client are built and gated; its connection to the corpus, a fragment pack from this project and a candidate generator that puts a second machine reading beside the first, is the next step. The arithmetic is worth stating: at roughly ten converged lines per player-hour it cannot cover 13.5 million lines, and it is not meant to; alignment reaches only the 2% of lines a printed edition covers, and on the other 98% a human reader is the only source of truth. The two methods occupy different slots, one improving the model everywhere, the other adjudicating the few thousand lines a year that matter most.

The Academy of Games is a digital institution in the spirit of Leibniz's *Drôle de Pensée*, with a Leibniz-in-character tutor that already cites a library of holdings. A search seam from that tutor to this project's API, carrying every citation with its line identifier, status and confidence, is designed and not yet built; the rule for it is already fixed: the tutor may say "the machine reads it as", never "Leibniz wrote".

8. Availability and reproduction

The repository (Apache-2.0) holds one Python package with one command-line tool, `leibniz`, whose subcommands run every stage from harvest to export; the reports are regenerated from the store by their own subcommands, and every number in this statement is one of them. Bulk data (the image cache, the store, the models) is never committed; the pipeline is cache-first and polite (one request per second per host, a User-Agent with a contact address, resumable everywhere). Tests run offline without the recognition stack. The companion deposits are: the corpus census (doi:10.5281/zenodo.22782815), the PHILIUMM reproduction and vision-language benchmark (doi:10.5281/zenodo.22782817), and the retro-alignment prototype, factory and audit (doi:10.5281/zenodo.22782819). This statement is versioned; later versions will record the second model's numbers and the release identifiers.

Acknowledgements

The images are the GWLB's, digitized and opened by the library; the catalogue is the BBAW's Leibniz-Katalog (TELOTA), CC BY 4.0; the recognition and segmentation models and the ground truth are the work of the ERC project PHILIUMM (Denisa-Florina Bumba, David Rabouin and colleagues, Laboratoire SPHERE, Université Paris Cité and CNRS), CC BY 4.0. The project builds on them and reports on them, nothing more.

References

- PHILIUMM HTR model for Leibniz's hand: doi:10.5281/zenodo.21457538; segmentation model: doi:10.5281/zenodo.21537859; HTR dataset for Leibniz's manuscripts: doi:10.5281/zenodo.21622297 (also huggingface.co/datasets/DenisaB/htr_leibniz_dataset_v1). CC BY 4.0.
- GWLB Digitale Sammlungen, https://digitale-sammlungen.gwlb.de/ (OAI-PMH, IIIF; Public Domain Mark 1.0, metadata CC0). Leibniz-Katalog, https://leibniz-katalog.bbaw.de/ (CC BY 4.0).
- Jungo, M., Vögtlin, L., Fakhari, A., Wegmann, N., Ingold, R., Fischer, A., Scius-Bertrand, A. (2023). Impact of Ground Truth Quality on Handwriting Recognition. SOICT 2023. arXiv:2312.09037.
- Peer, M., Scius-Bertrand, A., Fischer, A. (2025). CTC forced alignment for large-scale handwritten text recognition ground truth. arXiv:2508.07904.
- Peer, M., Scius-Bertrand, A., Scheurer, P., Fischer, A. (2026). BullingerDB: A Dataset for Handwritten Text Recognition and Writer Retrieval. arXiv:2605.30235.
- Tarride, S., Schneider, Y., Generali-Lince, M., Boillet, M., Abadie, B., Kermorvant, C. (2024). Improving Automatic Text Recognition with Language Models in the PyLaia Open-Source Library. arXiv:2404.18722.
- Camps, J.-B., Vidal-Gorène, C., Vernet, M. (2021). Handling heavily abbreviated manuscripts: HTR engines vs text normalisation approaches. arXiv:2107.03450.
- Clérice, T., Bawden, R., Glaise, A., Pinche, A., Smith, D. (2026). Pre-editorial normalization. arXiv:2602.13905.
- Boros, E., Ehrmann, M., Romanello, M., Najem-Meyer, S., Kaplan, F. (2024). Post-correction of historical text transcripts with large language models: an exploratory study. LaTeCH-CLfL 2024.
- Kanerva, J., Ledins, C., Käpyaho, S., Ginter, F. (2025). OCR error post-correction with LLMs in historical documents: no free lunches. arXiv:2502.01205.
- Humphries, M., et al. (2024). Unlocking the Archives: Using Large Language Models to Transcribe Handwritten Historical Documents. arXiv:2411.03340.
- Isom, S. (2025). An HTR-LLM workflow for high-accuracy transcription and analysis of abbreviated Latin court hand. arXiv:2507.04132.
- Muehlberger, G., et al. (2019). Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. Journal of Documentation 75(5), 954–976.
- van Strien, D., Beelen, K., Coll Ardanuy, M., Hosseini, K., McGillivray, B., Colavizza, G. (2020). Assessing the Impact of OCR Quality on Downstream NLP Tasks. ICAART 2020.
- Zhang, J., et al. (2025). OCR Hinders RAG: Evaluating the Cascading Impact of OCR on Retrieval-Augmented Generation. ICCV 2025. arXiv:2412.02592.
- MultiOCR-QA: Dataset for Evaluating Robustness of LLMs in Question Answering on Multilingual OCR Texts (2025). CIKM 2025. arXiv:2502.16781.
- Chrons, O., Sundell, S. (2011). Digitalkoot: Making Old Archives Accessible Using Crowdsourcing. HCOMP 2011.
- Leibniz Legible repository: https://github.com/marchofhares/leibnizlegible. Calculemus repository: https://github.com/marchofhares/calculemus. The Academy of Games repository: https://github.com/marchofhares/the-academy-of-games.