Reproducing the PHILIUMM Leibniz HTR model, and benchmarking frontier vision-language models on Leibniz's hand

Leibniz Legible, Phase B1

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

2026-09-16
doi:10.5281/zenodo.22782817 (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 report is: the first independent reproduction of the character error rate of the PHILIUMM project's handwritten-text-recognition model for Leibniz's hand, measured with a frozen, documented protocol on the model's own 1,878-line validation split (measured 7.95%, 95% CI 7.49 to 8.46, against the claimed 8.33%), together with the first published numbers for zero-shot frontier vision-language models on the same lines, which are four to ten times worse than the fine-tuned model. Two caveats the report states itself: the validation split is Latin and French by construction, so German and Kurrent are unmeasured; and the split is drawn from the dataset's manually corrected subset, so the number is a line-level figure on pre-segmented lines, not a page-level figure on raw scans. The model, segmentation model and ground truth are the work of the ERC project PHILIUMM (Denisa-Florina Bumba, David Rabouin and colleagues; Laboratoire SPHERE, Université Paris Cité and CNRS), released CC BY 4.0; this project builds on them and reports on them, nothing more.

PHILIUMM reproduction — HTR benchmark on the Leibniz val split (Phase B1)

Leibniz Legible, Phase B1 (gate). Generated 2026-09-04T20:49:31Z. The first independent reproduction of the PHILIUMM HTR model's character error rate, measured with a frozen, documented protocol before anything is built on it.

Gate verdict — ✅ REPRODUCED: measured CER 7.95% is within 1.0 point of the claimed 8.33% (Δ -0.38). Build on the PHILIUMM model.

Headline

Metric	Claimed (PHILIUMM)	Measured (this run)	Δ
CER	8.33%	7.95% (95% CI 7.49–8.46)	-0.38
WER	28.56%	27.04% (95% CI 25.95–28.19)	-1.52

Measured on **1,878 lines** of the PHILIUMM `val` split with the model's own recognition network (greedy CTC), under the **`philiumm`** normalization policy — the policy that mirrors how the model was *trained* (`normalization: NFD`, whitespace-collapsed), so this is an apples-to-apples comparison. (Re-scored from cached hypotheses; see the run record for inference time.)

Discrepancy analysis

- **Against the published 8.33% CER:** reproduced. Measured 7.95% vs claimed 8.33%.
- **Against the model's own metadata:** its `metadata.json` reports `accuracy: 0.9205` → a self-consistent character error of **7.95%** — our 7.95% tracks it (Δ +0.00). Note the model's own two self-reports (the 8.33% Zenodo headline and this 7.95% metadata figure) already differ by ~0.4 point — different val subsets or normalization — so there is no single canonical target; our protocol below fixes ours so the number is reproducible either way.
- **Why any gap at all:** the largest movable factor is normalization (see the sensitivity table); beyond that, line padding and the exact val subset each move CER a few tenths. None change the go/no-go.

Normalization sensitivity (why the policy is published, not assumed)

The same predictions score differently under different normalization. We report three policies; the headline uses `philiumm` (trained-with). Folding case and diacritics (`lenient`) only *lowers* CER and hides real errors, so it is shown for context, not as the headline.

Policy	Recipe	CER	WER
`philiumm`	NFD · collapse-ws · strip	7.95%	27.04%
`lenient`	NFKD · strip-diacritics · casefold · collapse-ws · strip	7.61%	25.40%

Policy	Recipe	CER	WER
`strict`	NFC · strip	7.95%	27.04%

Per-language breakdown

The GT ships **no per-line language labels** — the dataset features are only `text` + `image`, tagged `la, fr` at the dataset level (Latin + French). A line-level split therefore awaits the Phase C4 language-ID pass; reporting a guessed split here would be dishonest. This is the known German/Kurrent-gap caveat in miniature (SPECS §1.6): the number above is a **Latin+French** number by construction.

Error analysis — highest-CER lines

The worst lines are the diagnostic. Very short isolated lines (a single abbreviated word or number — often a marginal fragment) dominate the tail: with a tiny reference, one spurious extra word sends per-line CER **above 100%** (edit distance ÷ a 3-character reference). This is exactly why the corpus number is **micro-averaged** (length-weighted) — these fragments barely move it. Clean prose lines are frequently perfect. (Full per-line dumps: `reports/philiumm-repro.lines.jsonl`.)

Line	CER	Reference	Hypothesis
`val:00556`	333.33%	cum	cum nonnullis
`val:00455`	200.00%	Num	Hymluus
`val:00600`	200.00%	7	il
`val:00105`	177.78%	gradus differentiasque suas	quae precesserunt spos Deinde eaque tum quae …
`val:00194`	100.00%	quid dant 5 seu .	16
`val:01765`	100.00%	ab ipso	
`val:00245`	100.00%	licet	
`val:01837`	100.00%	primo	

464 of 1,878 lines (24.7%) are character-perfect under the `philiumm` policy; macro-averaged CER (per-line mean) is 9.16%.

Frontier-LLM comparison (zero-shot vision)

On a seeded random **150-line** subsample, the specialised HTR model and one or more frontier vision-language models were scored under the **identical protocol** — the first published LLM-on-Leibniz numbers:

Engine	CER	WER
Kraken (PHILIUMM, fine-tuned)	8.19% (95% CI 6.55–9.97)	26.86%
gpt-5.6-luna (zero-shot)	76.24% (95% CI 69.89–81.48)	90.89%
gpt-5.6-terra (zero-shot)	78.66% (95% CI 73.02–84.06)	91.08%
gpt-4o (zero-shot)	43.82% (95% CI 39.24–48.43)	76.58%
gpt-4.1 (zero-shot)	36.60% (95% CI 32.93–40.25)	70.72%
gpt-4.1-mini (zero-shot)	167.88% (95% CI 31.18–404.29)	262.55%
gemini-3.8-flash (zero-shot)	55.11% (95% CI 37.54–76.71)	88.94%

Engine	CER	WER
gpt-5.6-sol (zero-shot)	74.17% (95% CI 66.89–80.66)	84.39%

Token usage & estimated cost (from each API's `usage`; prices are approximate list rates, easily re-derived against current pricing):

Engine	Input tok	Output tok	Est. cost
gpt-5.6-luna	38,941	222,687	$0.275
gpt-5.6-terra	38,941	237,855	$2.932
gpt-4o	95,630	2,223	$0.261
gpt-4.1	95,630	2,202	$0.209
gpt-4.1-mini	48,094	6,309	$0.029
gemini-3.8-flash	172,399	2,936	$0.140
gpt-5.6-sol	38,941	241,151	$4.979

Zero-shot vision LLMs have never seen Leibniz's hand; the fine-tuned HTR model is expected to win decisively on secretary-hand Latin/French. The gap is the point — it quantifies how far a general model sits from a specialised one on this material, and it is far larger than the 8% CER the HTR model achieves.

What the artifacts contain

HTR model (Zenodo `10.5281/zenodo.21457538`, CC BY 4.0):

- `FoNDUE-GD_v2_ft_Leibniz.safetensors` — a Kraken `TorchVGSLModel` (safetensors container, `_kraken_min_version` 5.0.0), fine-tuned from `FoNDUE-GD_v2` (doi:10.5281/zenodo.14399779).
- Alphabet **178 graphemes** (Latn): Latin + French, plus Greek letters and a long tail of mathematical / astronomical symbols — a reminder the corpus is not pure prose.
- Ships its ketos training configs (`stage1_noisy.yml` / `stage2_clean.yml`) and the train/val file lists; the configs are the source of the `NFD` + whitespace normalization our default policy mirrors.

Ground truth (HuggingFace `DenisaB/htr_leibniz_dataset_v1`, CC BY 4.0):

Split	Lines
train_clean	18,254
train_noisy	43,372
val	1,878

- Rows are `(text, image)`; each image is an **already-extracted**, polygon-cropped and baseline-dewarped single line (verified by eye) — so the recogniser runs directly, no re-segmentation. `train_clean` is manually corrected; `train_noisy` is alignment-minted (Levenshtein ≥0.7 filter). The **`val` split is 1,878 lines**, drawn from the clean subset.
- **Segmentation model** (Zenodo `10.5281/zenodo.21537859`) is fetched for provenance but not needed here — we score pre-segmented lines (Phase B2/C1 will use it on raw pages).

Frozen protocol

Protocol `b1-2026-07` — pinned so numbers stay comparable:

- **version:** b1-2026-07
- **input:** pre-segmented single-line images paired 1:1 with reference text
- **default_policy:** philiumm
- **default_policy_detail:** NFD · collapse-ws · strip
- **aggregation:** micro-averaged (Σ edits ÷ Σ reference length)
- **distance:** Levenshtein, unit substitution cost, unicode-codepoint tokens
- **bootstrap:** 1000 resamples of lines with replacement, 95% percentile CI, seed=12345
- **notes:** Reference text is the GT as distributed; no manual correction. Engine output is transcribed once, deterministically where the engine allows. CER/WER preserve case and diacritics under the default policy (folding them only lowers the number and hides real errors).

Reproduce this

```
# 1. fetch artifacts (model + val split), cached under data/ (gitignored)
leibniz bench fetch
# 2. run the full reproduction (needs the kraken+torch stack; see below)
leibniz bench repro            # add --with-llm if ANTHROPIC_API_KEY is set
```

The kraken + torch stack is an **optional** dependency (`uv pip install kraken`), imported lazily: the harness, its metrics, and its tests all run without it. Kraken 7.0.3 loads the safetensors model via `kraken.models.loaders.load_models` ; inference uses `valid_norm=False` preprocessing (the baseline-model path) on the pre-extracted lines.

Gate

REPRODUCED: measured CER 7.95% is within 1.0 point of the claimed 8.33% (Δ -0.38). Build on the PHILIUMM model.