

The RISE Humanities Data Benchmark

From anecdote to evidence. Open, transparent, easy to re-use.

Maximilian Hindermann^{a, b} · Sorin Marti^a

^a RISE, University of Basel · ^b University Library Basel · maximilian.hindermann@unibas.ch

§01 Motivation

Every humanities project considering an LLM-based workflow faces the same question: *can this task be reliably automated?* Without a systematic answer, projects either skip LLMs entirely or invest in pipelines that fail late.

General-purpose benchmarks don't bridge the gap. They measure scale, not project-specific tasks. We replace anecdote with evidence: drop six files into a folder, open a pull request, and your task is benchmarked across every registered model.

§02 At a glance

AS OF MAY 2026

11 benchmark datasets	87 models	1,100+ test configurations	100K+ requests sent
--------------------------	--------------	-------------------------------	------------------------

§03 From folder to dashboard

THE WHOLE WORKFLOW

STEP 1 Fork \$ gh repo fork	STEP 2 Add files → benchmarks/<your_task>/	STEP 3 Run \$ python run_benchmarks.py	STEP 4 Open a PR \$ gh pr create	STEP 5 · AUTO Goes live ~ CI · results stream
-----------------------------------	--	--	--	---

— ↑ INSIDE STEP 02

images/ or texts/

page scans, cards, photos, text files, PDF files, ...



benchmark.py

scoring function

```
def score(pred, gt):
    return matches(pred, gt) \
        / len(gt)
```

dataclass.py

Pydantic schema

```
class Entry(BaseModel):
    name: str
    date: int
    city: str
```

ground_truths/

expert-verified outputs

```
{
  "name": "Alice",
  "date": 1923,
  "city": "Basel"
}
```

prompts/

prompt templates

```
Extract the listed fields
from the image as JSON.
Return only the JSON.
```

+ meta.json

task metadata

§04 What's in the suite

11 DATASETS

01 Bibliographic Data — metadata extracted from historical books	02 Blacklist Cards — structured fields from historical index cards
03 Book Advert XML — correcting malformed 18 th -c. advert XML	04 Business Letters — named entities from 20 th -c. Swiss letters
05 Company Lists — records from historical business directories	06 Fraktur Adverts — German Fraktur transcription, 16 th -20 th c.
07 General Meeting Minutes — agenda + votes from minute books	08 Library Cards — catalogue cards from historical libraries
09 Magazine Pages — advert bounding boxes on magazine pages	10 Medieval Manuscripts — 15 th -c. German, segmented + transcribed
11 Personnel Cards — 20 th -c. cards: position, salary, dates	

§05 Get involved

SCAN OR VISIT



CODE

github.com/rise-unibas/
humanities_data_benchmark



DASHBOARD

rise-services.rise.unibas.ch/benchmarks

§06 References and acknowledgements

[1] Hindermann, M., Kasper, L. K., Marti, S., & Bosse, A. (2026). From Experiments to Epistemic Practice: The RISE Humanities Data Benchmark. Journal of Open Humanities Data, 12(1), 38. doi:10.5334/johd.470

[2] Hindermann, M., Marti, S., Kasper, L. K., & Bosse, A. (2026). The RISE Humanities Data Benchmark: A Framework for Evaluating Large Language Models for Humanities Tasks. Journal of Open Humanities Data, 12(1), 24. doi:10.5334/johd.481

Thanks to the participating humanities projects at the University of Basel for sharing source material and ground truth.

CONTRIBUTORS A. Alberto · A. Binnenkade · A. Bosse · S. Burkhardt · E. Decker · P. Frick · M. Hindermann · L. Kasper · S. Lienhard · J. L. Losada Palenzuela · S. Marti · G. Müller · I. Serif · E. Spadini · T. Wullschlegler