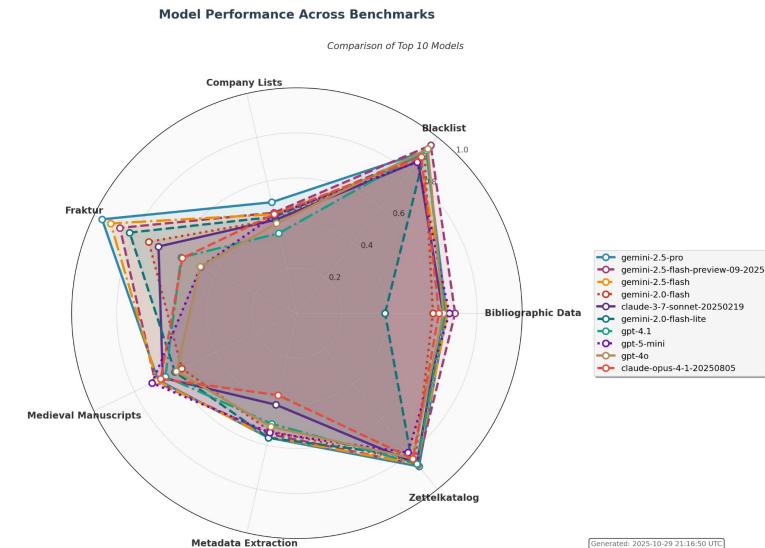


LLM-Benchmarking für die Digital Humanities

Ein praxisorientierter Ansatz aus der Forschungsberatung

Dr. Maximilian Hindermann (RISE & UB Basel) 11.11.2025



Programm

1. Benchmarking als epistemische Praxis
2. RISE Humanities Data Benchmark
3. Live Demo
4. Herausforderungen, Learnings und Ausblick
5. Fragen & Diskussion

Research and Infrastructure Support (RISE)



“Wir unterstützen Forscherinnen und Forscher in den Geistes- und Sozialwissenschaften an der Universität Basel bei Fragen der Konzeption computergestützter Forschung, der Herstellung, Analyse und nutzungsorientierten Darstellung von digitalen Daten, sowie deren nachhaltiger und offener Verfügbarmachung.”

<https://rise.unibas.ch>

Forschungsprojekte verwenden LLMs

Männliche Angestellte.		Swiss TPH Swiss Tropical and Public Health Institute
Name und Vorname:	Baumgartner Gottfried	
Geburtsdatum:	8.8.1900	Eremit
Heimatort:	Basel und Sirnach/Thurgau	
Letzter Eintritt in den Dienstinstanz:		1. Mai 1944
*) Dienstjahr zählen ab:		
*) Spareinleger seit:		
*) Versichert seit:		
*) Versicherungsjahre: wird vom Personalamt ausgefüllt		
Bildungsgang und Tätigkeit		
4 Jahre Primar-, 4 Jahre Unter- schule Basel, 3-jährige kaufmän-		
Fremdsprachkorrespondent in Bas- ausländischer Fabrikationsfirma		
Bemerkungen:	Zivilstand: verheiratet Kinder unter 18 Jahren: zivilrecht. Wohnsitz: B. Milit. Einteilung: Pol.S	
Form. Pa. Nr. 2 - IX. 42 - 10 000 - 63115		

Swiss TPH 
Public and Health Institute

Reisemedizin Studienangebot Forschung Services

Projects > LUDOK

Erweiterte Suche

Schlagwort	Autor(en)	Titel
Orte	Studientyp	Zielgruppe
Reaktionsbereitschaft	Experimentelle Studien Epidemiologische Studien Übersichten, Methodik	Kinder Erwachsene
salvoziale Lavage		
Unfall		
Kohlenwasserstoffe		
Bronchitis		
offe		
Job, Partikel	Region	Kollektiv
Oxidantien	Europa Übrige Länder	Säuglinge, Vorschulkinder Schulkinder und Jugendliche Erwachsene (alle)
Ammoniak		Betagierte Personen (65+) Personen mit Asthma oder Pers. mit anderer chron.
Wasserstoff, nicht chloriert, VOC		Registeraten, Patienten Schwangere Frauen
NOx		
Schwefelverbindungen,		
halogenierte Stoffe		
Typ	Spezies	Zeitdauer
Entleinte Studie	Mensch Tier	Kurzfristig Langfristig
Exposition, Unfall, Brand		
Panel, Kurzfristige Längssstudien	Viren, Bakterien, Einzeller	
init, Fall-Kontrollstudie, deskriptiv		
Studie		
Studiemethodik		

<https://owlabs.com/products/meeting-q3-2024-10-29>

Seite 1 bis 20 von 22



Warum Benchmarking?

Epistemischer Grund:

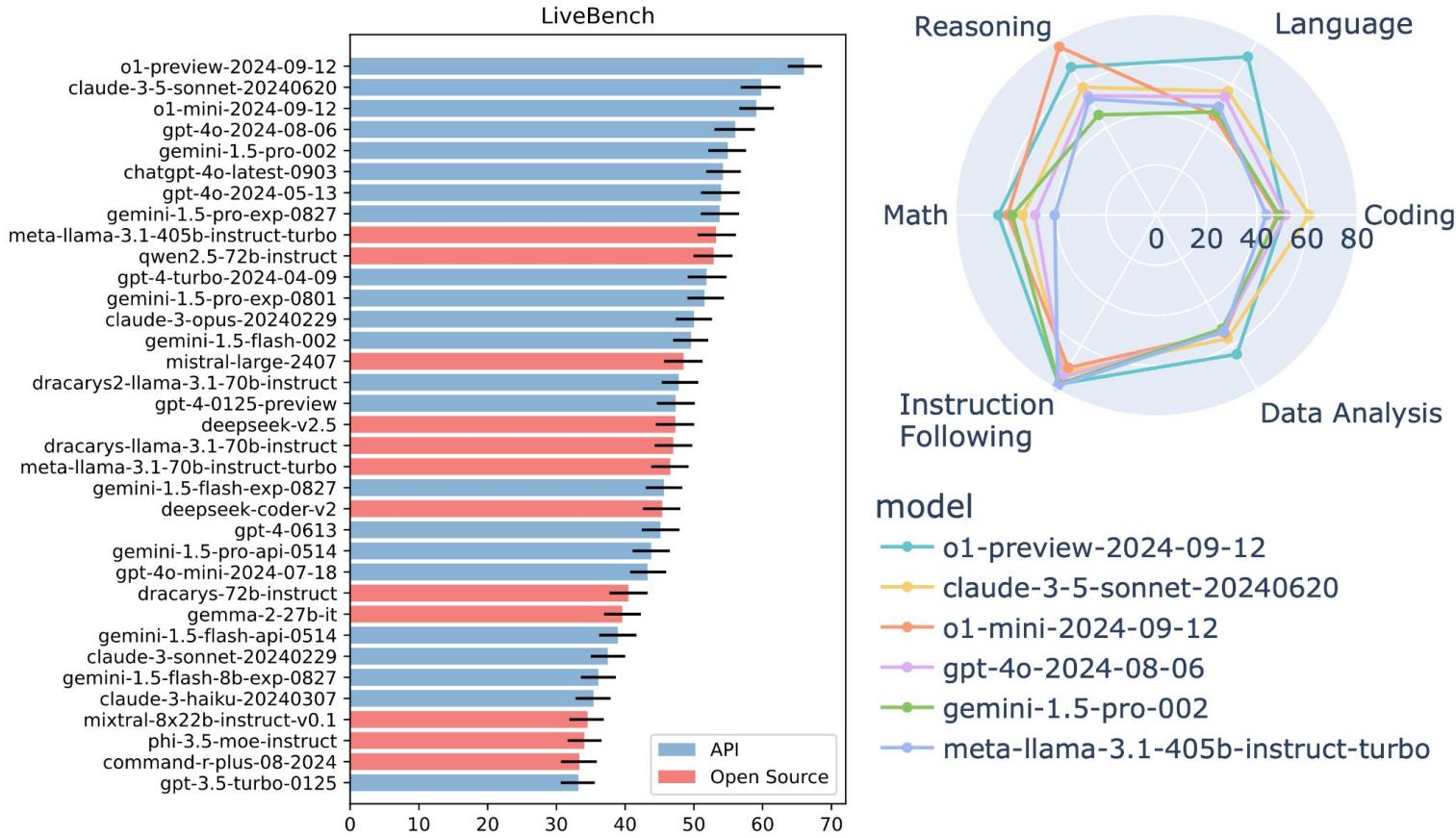
- Evidenzbasierte Entscheidungsfindung darüber, welche(s) Modell(e) für welche Aufgabe(n) in einem Forschungsprojekt eingesetzt werden sollen.

Praktische Gründe. Zwingt Forschungsprojekte:

- Aufgaben zu operationalisieren
- Festzulegen, was “gut genug” ist
- Fertigkeiten und Budget gemäss Umsetzbarkeit zu prüfen
- Rechtliche oder ethische Fragen zu klären

Was ist Benchmarking?

Ein Modell bzw. eine Menge von Modellen für eine Aufgabe bzw. eine Menge von Aufgaben anhand eines Goldstandards und einer Metrik pro Aufgabe einstufen oder bewerten.



[Colin White et al. \(2024\). LiveBench: A challenging, contamination-free LLM benchmark. arXiv preprint arXiv:2406.19314.](#)

question_id	category	ground_truth	turns	task
string · lengths	string · classes	string · lengths	sequence · lengths	string · classes
64	reasoning	1+3	1	spatial
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fc1191f1dc0bb	reasoning	3	["Suppose I have a physical, solid, equilateral triangle, and I make two cuts. The two cuts are from two parallel lines, and both cuts pass through the interior of the triangle. How many pieces are there after the cuts? Think step by step, and then put your answer in **bold** as a single integer (for example, **0**). If you don't know, guess."]	spatial
95ef46b58759a1595006301bebb7be5edff15ad5fcac453d3483c6d8d554b10	reasoning	2	["Suppose I have a physical, solid,..."	spatial
efb9d081389f7fe028eb9a801d6751ab5cc6a2f4d09e21e1ca6c893e9b8f2fc5	reasoning	1	["Suppose I have a physical, solid,..."	spatial
3d118e7c265b1ddf3c22d8765dc0c224dec03fad27d51c9687b46d61b6c4d0e5	reasoning	4	["Suppose I have a physical, solid square wit..."	spatial

<https://huggingface.co/datasets/livebench/reasoning>, 2024-10-28

question_id	answer_id	model_id	choices
string · lengths 64 100%	string · lengths 22 100%	string · classes o1-mini-20... 1.1%	list · lengths 1 100%
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	NvEWL5ecvoYrKM64PT9bb8	Phi-3-small-8k-instruct	[{ "index": 0, "turns": ["***4**"] }]
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	7iFSHBmedsNoVBmLNjQqW6	Qwen1.5-4B-Chat	[{ "index": 0, "turns": ["There will be 3 pieces after the cuts."] }]
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	PRhP3nkqS3a2rMSWt8nDeQ	o1-mini-2024-09-12	[{ "index": 0, "turns": ["***3**"] }]
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	DP2nByzvdSXRBdg9ZUk3Wg	Qwen2-1.5B-Instruct	[{ "index": 0, "turns": ["There are 4 pieces."] }]
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	c44auEizcaZzR466m2AiXH	Qwen1.5-0.5B-Chat	[{ "index": 0, "turns": ["The number of pieces that will be created is $2^3 = 8$."] }]
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	c9NdbqeA8Unj8rFSyCLHez	Qwen2-0.5B-Instruct	[{ "index": 0, "turns": ["The number of pieces is 3."] }]
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	6VNYYfpPvXUrnKk6KDQao2b	gpt-3.5-turbo-1106	[{ "index": 0, "turns": ["Sure, let's think through this step by step. \\n\\nWhe..."] }]
59979242f2437bc604ba87e9af3dd80878837a3ad01436d110fca1191f1dc0bb	Nfwp5caHCxEPBygxTpDf4c	gpt-4-0613	[{ "index": 0, "turns": ["If you make two cuts through the interior of the..."] }]

https://huggingface.co/datasets/livebench/model_answer, 2024-10-28

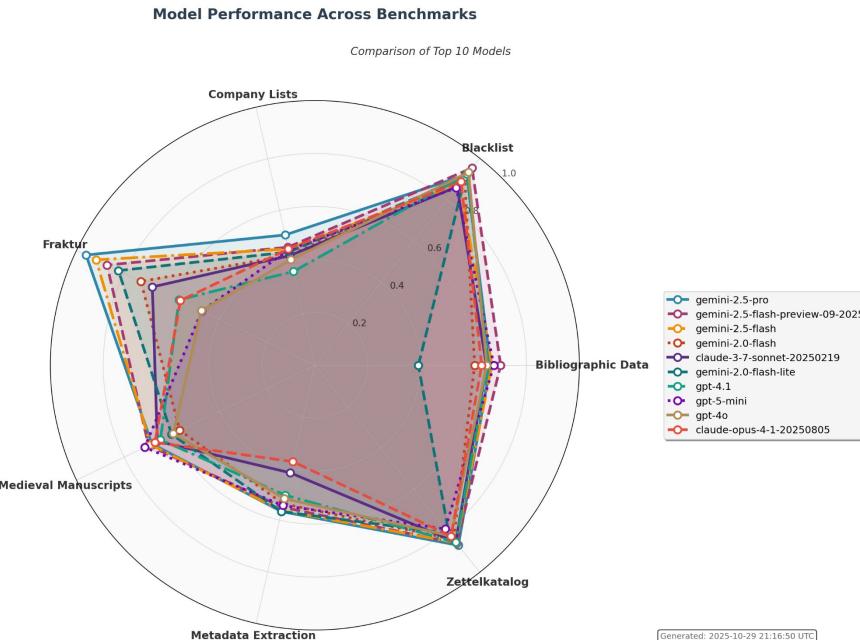
Problemstellung

- Viele Beratungsanfragen mit ähnlichen Desiderata bezüglich LLMs:
 - Segmentierung und Transkription
 - (Meta)datenextraktion
- Viele gute Benchmarks, aber: nicht spezifisch genug für vorliegendes geisteswissenschaftliches Material
- Grosser Aufwand, Projekte isoliert zu beraten resp. zu benchmarken (verteiltes Wissen und verteilte Infrastruktur)

Lösung

- Einzelne Benchmarks standardisieren und zusammenführen: Vergleichbar, reproduzierbar, erweiterbar
- Gemeinsame Infrastruktur statt Einzelfalllösungen
- Benchmarking als fester Bestandteil der Projektberatung
- Forschungsprojekte befähigen, eigene Benchmarks umzusetzen

RISE Humanities Data Benchmark



gemini-2.5-flash-preview-09_2025	Google	T0219	2025-10-01	prompt.txt	None	Fuzzy	0.702	\$0.0307	\$0.0437	116.65	166.10
gpt-5	OpenAI	T0129	2025-10-01	prompt.txt	None	Fuzzy	0.685	\$0.3421	\$0.4992	591.90	863.65
gpt-5-mini	OpenAI	T0130	2025-10-01	prompt.txt	None	Fuzzy	0.677	\$0.0582	\$0.0860	411.12	607.56
gemini-2.5-flash	Google	T0195	2025-09-30	prompt.txt	None	Fuzzy	0.666	\$0.0252	\$0.0376	195.82	292.59
claude-connet-4_20290514	Anthropic	T0107	2025-09-30	prompt.txt	None	Fuzzy	0.669	\$0.1692	\$0.2531	127.79	191.16
o3	OpenAI	T0131	2025-10-01	prompt.txt	None	Fuzzy	0.667	\$0.1885	\$0.2827	391.04	586.48
gemini-2.5-pro	Google	T0128	2025-09-30	prompt.txt	None	Fuzzy	0.664	\$0.1032	\$0.1554	227.18	342.25
mistral-medium-2505	Mistral AI	T0126	2025-10-01	prompt.txt	None	Fuzzy	0.661	\$0.0222	\$0.0336	128.32	194.01
gpt-4.1	OpenAI	T0139	2025-10-01	prompt.txt	None	Fuzzy	0.657	\$0.0952	\$0.1449	298.94	455.07
qwen/qwen3-v1-fb-thinking	Alibaba (via OpenRouter)	T0223	2025-10-17	prompt.txt	None	Fuzzy	0.657	\$0.1268	\$0.1931	923.12	1405.84
mistral-medium-2508	Mistral AI	T0166	2025-10-01	prompt.txt	None	Fuzzy	0.654	\$0.0220	\$0.0336	112.70	172.31
claude-3.5-connet-20241022	Anthropic	T0069	2025-09-30	prompt.txt	None	Fuzzy	0.651	\$0.1682	\$0.2576	124.19	190.17
gpt-4.0	OpenAI	T0067	2025-09-30	prompt.txt	None	Fuzzy	0.650	\$0.1136	\$0.1748	350.22	538.95
claude-5.7-connet-20290719	Anthropic	T0031	2025-09-30	prompt.txt	None	Fuzzy	0.649	\$0.1765	\$0.2720	136.48	210.38
gpt-4.1-mini	OpenAI	T0140	2025-10-01	prompt.txt	None	Fuzzy	0.646	\$0.0199	\$0.0307	164.93	254.41
mistral-large-latest	Mistral AI	T0187	2025-10-01	prompt.txt	None	Fuzzy	0.639	\$0.0805	\$0.1259	136.28	213.12
claude-cpus-4-1-20290605	Anthropic	T0127	2025-09-30	prompt.txt	None	Fuzzy	0.631	\$0.9735	\$1.5435	203.32	322.38
meta-flame/flame-4-maverick	Meta (via OpenRouter)	T0234	2025-10-17	prompt.txt	None	Fuzzy	0.630	\$0.0062	\$0.0099	151.02	241.35
gemini-2.0-flash	Google	T0068	2025-09-30	prompt.txt	None	Fuzzy	0.634	\$0.0052	\$0.0087	69.66	115.32
gpt-5-mini	OpenAI	T0131	2025-10-01	prompt.txt	None	Fuzzy	0.570	\$0.0281	\$0.0476	401.62	681.07
claude-cpus-4-20290514	Anthropic	T0106	2025-09-30	prompt.txt	None	Fuzzy	0.561	\$0.8992	\$1.5413	193.49	331.67
gemini-2.5-flash-line-preview-09_2025	Google	T0211	2025-10-01	prompt.txt	None	Fuzzy	0.579	\$0.0048	\$0.0083	18.69	32.28
gemini-2.5-flashlite	Google	T0203	2025-10-01	prompt.txt	None	Fuzzy	0.545	\$0.0039	\$0.0072	19.33	35.50

[Hindermann, M., Marti, S., Alberto, A., Burkhardt, S., Decker, E., Frick, P., Kasper, L., Losada Palenzuela, J. L., Müller, G., Serif, I., & Spadini, E. \(2025\). RISE-UNIBAS/humanities_data_benchmark \(v0.3.1\). Zenodo.](https://doi.org/10.5281/zenodo.1747519)
<https://doi.org/10.5281/zenodo.1747519>

RISE HDB: Factsheet

- Modulare Benchmarks (7 live und 10+ in Vorbereitung)
- Monatliche Testläufe (400+ Tests mit jeweils mind. 1 Lauf)
- Bewertung pro Benchmark pro Modell inkl. Kosten und Zeit
- Normalisierte Bewertung zwischen Benchmarks inkl. Kosten und Zeit
- Spezialisiert auf structured outputs von kommerziellen multimodalen LLMs
- Preisbeobachtung mittels Wayback Machine Snapshots
- API agnostisch (mittels <https://pypi.org/project/generic-llm-api-client/>)
- FAIR Software und Daten (Benchmarks und Resultate)
- Automatisch bespieltes Frontend zur einfacheren Sichtung

RISE HDB Live Demo: Neuen Benchmark erstellen

- Ordner in benchmarks befüllen:

```
benchmarks/<benchmark_name>/  
├── README.md          (use README_TEMPLATE.md)  
├── images/             (image files: jpg, png)  
├── prompts/            (text files with prompts)  
├── ground_truths/      (json or txt files)  
└── benchmark.py        (custom scoring)  
    └── dataclass.py     (optional: Pydantic models for structured output)
```

- Eintrag (“Test”) für jede gewünschte Konfiguration in benchmark_tests.csv.

RISE HDB Live Demo: aktuelles Frontend (v0.3.1)

Humanities Data Benchmark

Welcome to the **Humanities Data Benchmark** report page. This page provides an overview of all benchmark tests, results, and comparisons.

Leaderboard

The table below shows the **global average performance, cost efficiency, and time efficiency** of each model across the seven core benchmarks: bibliographic_data, blacklist, company, lists, flickr, medieval_ms_transcripts, metadata_extracts, and metathesaurus.

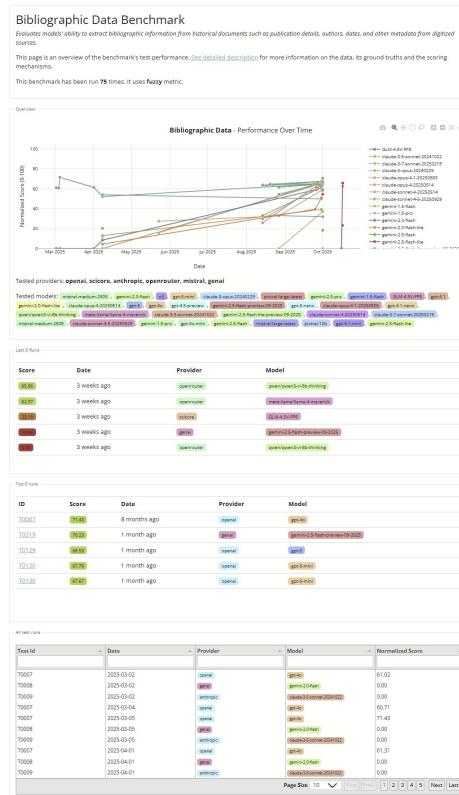
The **Model** and **Provider** columns identify each system. **Global Average** represents the mean performance score across all seven benchmarks (higher is better). **CostPoint** and **TimePoint** show normalized efficiency metrics calculated per test, averaged per provider, then averaged globally; this multi-level normalization accounts for differences of items, test configurations, and benchmark results. For efficiency metrics, lower values are better. Individual tests cost or time needed per performance point achieved.

In the seven benchmark specific columns show average performance for each individual benchmark. Only models with results in all seven benchmarks are included. Click on any column header to sort the table.

Model	Provider	Global Average	CostPoint	TimePoint	bibliographic_data	blacklist	company	flickr	medieval_ms_transcripts	metadata_extracts	metathesaurus
T001	Global Average	0.799	0.2347	37.899	green	green	green	green	green	green	green
T002	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T003	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T004	Global Average	0.799	0.2347	37.899	green	green	green	green	green	green	green
T005	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T006	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T007	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T008	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T009	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T010	Global Average	0.798	0.2348	37.899	green	green	green	green	green	green	green
T011	Global Average	0.699	0.0249	11.890	green	green	green	green	green	green	green
T012	Global Average	0.690	0.0180	24.770	green	green	green	green	green	green	green
T013	Global Average	0.649	0.0197	10.050	green	green	green	green	green	green	green
T014	Global Average	0.649	0.0197	10.050	green	green	green	green	green	green	green
T015	Global Average	0.649	0.0197	10.050	green	green	green	green	green	green	green
T016	Global Average	0.649	0.0197	10.050	green	green	green	green	green	green	green
T017	Global Average	0.643	0.0305	82.076	green	green	green	green	green	green	green
T018	Global Average	0.643	0.0305	82.076	green	green	green	green	green	green	green
T019	Global Average	0.629	0.0174	63.480	green	green	green	green	green	green	green
T020	Global Average	0.629	0.0174	63.480	green	green	green	green	green	green	green
T021	Global Average	0.629	0.0174	63.480	green	green	green	green	green	green	green
T022	Global Average	0.617	0.0302	28.290	green	green	green	green	green	green	green
T023	Global Average	0.605	0.0340	34.230	green	green	green	green	green	green	green
T024	Global Average	0.605	0.0340	34.230	green	green	green	green	green	green	green
T025	Global Average	0.603	0.0302	24.020	green	green	green	green	green	green	green
T026	Global Average	0.603	0.0302	24.020	green	green	green	green	green	green	green
T027	Global Average	0.597	0.0396	87.870	green	green	green	green	green	green	green
T028	Global Average	0.597	0.0396	87.870	green	green	green	green	green	green	green
T029	Global Average	0.596	0.0372	79.950	green	green	green	green	green	green	green
T030	Global Average	0.595	0.0372	79.950	green	green	green	green	green	green	green
T031	Global Average	0.595	0.0372	79.950	green	green	green	green	green	green	green
T032	Global Average	0.595	0.0372	79.950	green	green	green	green	green	green	green
T033	Global Average	0.594	0.0305	130.000	green	green	green	green	green	green	green
T034	Global Average	0.594	0.0305	130.000	green	green	green	green	green	green	green
T035	Global Average	0.588	0.0378	5.450	green	green	green	green	green	green	green
T036	Global Average	0.588	0.0378	5.450	green	green	green	green	green	green	green
T037	Global Average	0.588	0.0378	5.450	green	green	green	green	green	green	green
T038	Global Average	0.588	0.0378	5.450	green	green	green	green	green	green	green
T039	Global Average	0.579	0.0356	11.400	green	green	green	green	green	green	green
T040	Global Average	0.554	0.0302	42.710	green	green	green	green	green	green	green
T041	Global Average	0.554	0.0302	42.710	green	green	green	green	green	green	green
T042	Global Average	0.544	0.0192	44.800	green	green	green	green	green	green	green
T043	Global Average	0.542	0.0191	43.180	green	green	green	green	green	green	green
T044	Global Average	0.540	0.0191	43.180	green	green	green	green	green	green	green
T045	Global Average	0.540	0.0191	43.180	green	green	green	green	green	green	green
T046	Global Average	0.540	0.0191	43.180	green	green	green	green	green	green	green
T047	Global Average	0.493	0.1710	19.490	green	green	green	green	green	green	green
T048	Global Average	0.493	0.1710	19.490	green	green	green	green	green	green	green
T049	Global Average	0.493	0.1710	19.490	green	green	green	green	green	green	green

- https://rise-unibas.github.io/humanities_data_benchmark/
- GitHub Pages Instanz
- Seitenrenderings pro Benchmark
- Keine Datenbank

RISE HDB Live Demo: neues Frontend (geplant v0.4.0)



- Django-App basierend auf NDR-Core
 - MongoDB
 - Umfangreiche Suchen
 - Datenvisualisierung mit Plotly

Herausforderungen

- Hoher Aufwand bei Ground Truth Erstellung wird systematisch unterschätzt
 - Weiterverarbeitung generierter Daten oft unklar
 - Hohe interpretative Dichte
 - Feld extrem dynamisch
 - Kosten
-
- ✓ Pilotprojekte mit echten Daten
 - ✓ MosAlc, Nodegoat, etc.
 - ✓ Trennung von Interpretation und Messung
 - ✓ Mut zur Lücke
 - ✓ EDA, Beratungshonorare, Dritt- und Zweitmittel

Learnings

- Benchmarking als epistemische Praxis:
 - Explizite Interpretationsentscheidungen
 - Projektspezifische Erfolgskriterien
 - Transparente Methodendokumentation
 - Fokus auf operationalisierbare Komponenten
- Modelle unterscheiden sich deutlich in Leistung, Kosten und Robustheit je nach Aufgabe.

Ausblick

- Finanzierung absichern
- Governance Onboarding externer Benchmarks
- Erweiterung auf
 - neue Datentypen (besonders Text)
 - neue Klassen von Aufgaben (besonders Anreicherung)
 - spezialisierte Modelle (besonders ATR)
- Bessere Integration mit sciCORE (HPC Universität Basel)
- Langzeitbeobachtung von Modellleistung

RISE HDB: Mitwirkende



Gabriel Müller
Domain Expert, Data Curator,
Annotator



Sven Burkhardt
Annotator



Maximilian Hindermann
Data Curator, Annotator, Analyst,
Engineer



Pema Frick
Domain Expert, Data Curator,
Annotator, Analyst, Engineer



Anthea Alberto
Data Curator, Annotator



Elena Spadini
Data Curator, Annotator



Sorin Marti
Data Curator, Annotator, Analyst,
Engineer



Ina Serif
Domain Expert, Data Curator,
Annotator



Lea Kasper
Domain Expert, Data Curator,
Annotator, Analyst, Engineer



Eric Decker
Data Curator, Annotator



José Luis Losada
Palenzuela
Data Curator, Annotator

Mitmachen & Kontakt

RISE Humanities Data Benchmark

https://github.com/RISE-UNIBAS/humanities_data_benchmark

<https://doi.org/10.5281/zenodo.16941752>

Entwicklerteam

Maximilian Hindermann

maximilian.hindermann@unibas.ch

<https://orcid.org/0000-0002-9337-4655>

Sorin Marti

sorin.marti@unibas.ch

<https://orcid.org/0000-0002-9541-1202>

RISE

<https://rise.unibas.ch>

<https://github.com/RISE-UNIBAS/>

Fragen und Diskussion

