

famulus-logic

Ein Prüfstand für Klassifikatoren logischer Beziehungen zwischen Aussagen der PDAI-Sammlungen

2026-10-11

■ Inhaltsverzeichnis

1 Zweck	1
2 Grundbegriffe	2
3 Der Relationskatalog	2
3.1 Das Vier-Zellen-Verfahren	2
3.2 Die sieben Labels	3
3.3 Wie man eine Relation feststellt	4
3.4 Beispiele des Katalogs	5
3.5 Gründe einer Relation	5
3.6 Beinahe-Treffer	6
3.7 Zulassung als Aussage	7
4 Aufbau	8
5 Ernte der Sätze	8
5.1 Quellen	8
5.2 Satzteilung und Vorfilter	9
5.3 Ernte des Piloten	9
6 Erzeugung der Proben	10
6.1 Verfahren	10
6.2 Prüfung und Gold-Status	10
6.3 Benchmark des Piloten	10
7 Benchmark-Aufgaben	12
8 Klassifikatoren	13
9 Kennzahlen	13
10 Die Lauftabelle	14
11 Ergebnisse des Piloten	14
11.1 Läufe	14
11.2 Rangliste	15
11.3 Fehlermatrizen der aktuellen LLM-Läufe	16
11.4 Lexikalische Basislinie	18
11.5 Beinahe-Treffer nach Typ	19
11.6 Genauigkeit nach Grund	20
11.7 Vergleich der Modelle	20
12 Prüfung des Goldes	22
13 Bedienung	22
14 Grenzen und nächste Schritte	23

■ 1 Zweck

famulus-logic misst, wie gut ein Klassifikator entscheidet, ob eine **Probe** — ein vorgegebener Ausdruck P — zu irgendeinem Satz einer Liste von Korpussätzen $S_1 \dots S_n$ in einer logischen Beziehung steht, und in welcher. Die Korpussätze stammen aus den PDAI-Sammlungen (Briefe, Editionen, Sekundärliteratur); die Proben werden zu jedem Satz erzeugt, und zwar so, dass sie die Beziehung entweder wirklich haben oder ihr nur ähneln (**Beinahe-Treffer**).

Der Prüfstand liefert drei Dinge:

1. einen **Relationskatalog** (`config/relations.json`), der jede Beziehung definiert und beschreibt, wie man sie feststellt — für Menschen und als Anweisung an den Klassifikator;
2. eine **Satztable** (`data/benchmark.json`): Aussagen aus den Sammlungen mit Herkunft und, als Attribute, den Proben samt Gold-Label;
3. eine **Lauf**table (`runs/bench.duckdb`, Tabelle `runs`): je Klassifikatorlauf die Fehlermatrix, die Kennzahlen, Wandzeit und Durchsatz sowie die Eigenschaften der benutzten Klassifikator-Version.

Die Folgerung, um die es geht, ist **nicht die materiale Implikation**. Ein Satz folgt aus einem anderen auch kraft Begriffsregeln (*Junggeselle* $\rightarrow$ *unverheirateter Mann*), Synonymie, Unterbegriffen, konversen Relationen, Kalenderrechnung oder Ortseinschluss — sofern die Regel eine Regel der Bedeutung ist und keine unausgesprochene Tatsache.

2 Grundbegriffe

Aussage

Ein Satz, der wahr oder falsch ist. Nur Aussagen werden zugelassen, als Korpussatz wie als Probe. Fragen, Anreden, Überschriften, Literaturangaben und Seitenbruch-Fragmente sind keine Aussagen.

Korpussatz S

Wörtlich aus einer PDAI-Seite übernommen, mit historischer Schreibung; nur Leerraum wird normalisiert.

Probe P

Ein selbständiger Aussagesatz, vom Generator zu genau einem Korpussatz geschrieben. Er erbt dessen Kontext (Personen, Daten, Orte).

Relation

Die Beziehung **von S zu P**. `entails` heißt $S \models P$, `entailed_by` heißt $P \models S$.

Beinahe-Treffer

Eine Probe, die aussieht, als hätte sie die Relation `mimics` zu S, deren wahre Relation (Gold) aber eine andere ist.

Aufgabe

Eine Probe mit einer Kandidatenliste: ihr eigener Satz (Quellpaar) und Ablenker aus anderen Dokumenten und aus demselben Dokument.

3 Der Relationskatalog

3.1 Das Vier-Zellen-Verfahren

Jede Relation ist dadurch definiert, welche Kombinationen von Wahrheitswerten des Paares (S, P) **unmöglich** sind. Der Klassifikator beurteilt die vier Zellen; das Label folgt daraus eindeutig.

Zelle	Bedeutung
TT	S wahr und P wahr
TF	S wahr und P falsch
FT	S falsch und P wahr

Zelle	Bedeutung
FF	S falsch und P falsch

Das Verfahren in sieben Schritten (wörtlich aus dem Katalog):

1. S und P als Aussagen lesen; Namen, Daten und Indexikalien so festlegen, wie S sie festlegt — P erbt diesen Kontext.
2. TF: Kann S wahr sein, während P falsch ist? Wenn nein, folgt P aus S.
3. FT: Kann P wahr sein, während S falsch ist? Wenn nein, folgt S aus P.
4. TT: Können beide zugleich wahr sein? Wenn nein, sind sie unverträglich.
5. FF: Können beide zugleich falsch sein? Wenn nein, gilt mindestens eine.
6. Die unmöglichen Zellen auf das Label abbilden. Braucht man eine unausgesprochene Tatsachenprämisse, um eine Zelle auszuschließen, bleibt sie möglich — es sei denn, die Prämisse ist eine Begriffsregel.
7. Thematische Ähnlichkeit, gemeinsame Wörter oder Namen sind für sich nie ein Beleg für eine Relation.

Weil zwei Labels nie dieselbe Menge unmöglicher Zellen haben, prüft der Katalog beim Laden seine eigene Eindeutigkeit. Die Abbildung erlaubt zwei Entscheidungsarten des Klassifikators: das Label aus den Zellen ableiten (`decide=cells`) oder dem genannten Label trauen (`decide=label`). Die Abweichung zwischen beiden ist selbst eine Qualitätsangabe.

3.2 Die sieben Labels

Label	Name	un-möglich	Famili	Bedeutung
<code>equivalent</code>	Äquivalenz	TF, FT	follows	S and P are true in exactly the same situations: each follows from the other. Paraphrase, synonym substitution, active/passive, converse relational terms and definitional unfolding produce equivalents.
<code>entails</code>	Folgerung ($S \models P$)	TF	follows	P follows from S, but S does not follow from P: P is weaker. Typical: hypernym in an upward-monotone position, dropping a conjunct or a restrictive modifier in an upward context, existential generalisation, a conceptual rule $A \rightarrow B$, a date that is coarsened.
<code>entailed_by</code>	Rückfolgerung ($P \models S$)	FT	follows	S follows from P, but P does not follow from S: P is stronger or more specific than S.
<code>contradictory</code>	Kontradiktion	TT, FF	opposes	Exactly one of S and P is true: P is equivalent to the negation of S. Complementary antonyms (lebendig/tot, verheiratet/unverheiratet for persons) and wide-scope negation produce contradictions.
<code>contrary</code>	Konträrität	TT	opposes	S and P cannot both be true, but both can be false. Incompatible alternatives from a range with more than two members: different places or dates for

Label	Name	un- mög- lich	Fami- lie	Bedeutung
				the same singular event, scalar antonyms (heiß/kalt), ‚alle‘ vs. ‚keine‘.
subcontrary	Subkonträrität	FF	opposes	S and P cannot both be false, but both can be true: at least one holds. Typical: ‚einige sind F‘ vs. ‚einige sind nicht F‘ (with existential import), or P being the negation of something S entails’ contrary.
independent	Unabhängigkeit	—	none	All four truth-value combinations are possible: neither follows from the other and they are compatible. This is the correct label for near misses that only look related and for distractors from other documents.

Für die grobe Fehlermatrix werden die Labels zu drei Familien zusammengefasst: `follows` (eine Richtung der Folgerung), `opposes` (Unverträglichkeit oder Erschöpfung), `none`.

3.3 Wie man eine Relation feststellt

Diese Texte gehen unverändert in die Anweisung an den Klassifikator ein.

Label	Feststellung (Anweisung an den Klassifikator)
equivalent	Check both directions separately. $S \models P$ and $P \models S$ must both hold. If P adds any detail (a date, a qualifier, a stronger quantifier) or drops any detail of S, it is not equivalent but entails/entailed_by.
entails	Find the rule that carries S to P (state it). Then confirm the converse fails: construct a situation where P is true and S false. Watch monotonicity: under negation or ‚kein/alle-Restriktor‘ the direction reverses.
entailed_by	Mirror of ‚entails‘: the rule carries P to S, and a situation exists where S is true while P is false.
contradictory	Confirm both: they cannot be true together AND they cannot be false together. If a third possibility makes both false (another place, another date, a middle value), the pair is merely contrary.
contrary	Find the third possibility that makes both false. If none exists under the conceptual rules, the pair is contradictory.
subcontrary	Show that the falsity of S forces the truth of P, and find a situation in which both are true. Rare in corpus sentences; generate only where natural.
independent	For each candidate rule that might connect S and P, test whether it is a conceptual rule or a factual premise. If only a factual premise (or wishful reading) connects them, the pair is independent.

3.4 Beispiele des Katalogs

Label	S	P	Grund
equivalent	Er ist Junggeselle.	Er ist ein unverheirateter Mann.	conceptual_rule
equivalent	Kepler schrieb an Maestlin.	Maestlin war Empfänger eines Briefes von Kepler.	converse_relation
entails	Er ist Junggeselle.	Er ist unverheiratet.	conceptual_rule
entails	Der Brief wurde im März 1598 in Graz geschrieben.	Der Brief wurde 1598 geschrieben.	numeric_temporal
entails	Tycho beobachtete den Mars.	Tycho beobachtete einen Planeten.	hyponymy
entailed_by	Er ist unverheiratet.	Er ist Junggeselle.	conceptual_rule
entailed_by	Tycho beobachtete einen Planeten.	Tycho beobachtete den Mars.	hyponymy
contradictory	Er ist Junggeselle.	Es ist nicht der Fall, dass er Junggeselle ist.	formal
contradictory	Alle Planeten bewegen sich um die Sonne.	Einige Planeten bewegen sich nicht um die Sonne.	formal
contrary	Er ist Junggeselle.	Er ist mit Barbara verheiratet.	conceptual_rule
contrary	Der Brief wurde in Graz geschrieben.	Der Brief wurde in Prag geschrieben.	part_whole_place
subcontrary	Einige Beobachtungen Tychos sind genau.	Einige Beobachtungen Tychos sind nicht genau.	formal
independent	Er ist Junggeselle.	Er ist ein junger Mann.	lexical_overlap
independent	Kepler glaubte, dass die Bahn des Mars elliptisch sei.	Die Bahn des Mars ist elliptisch.	attitude_context

3.5 Gründe einer Relation

Jede Probe trägt den Grund (basis), auf dem ihre Relation beruht, und — für Begriffsregeln, Synonymie, Unterbegriffe, konverse Relationen, Zahl/Zeit, Ort und Weltwissen — die Regel in der Form $A \rightarrow B$. Der Grund ist kein Label, aber er erlaubt die Fehleranalyse nach der Art der Folgerung.

Grund	Beschreibung	Prüfung
formal	Logical form alone: negation, conjunction elimination, quantifier laws, square of opposition.	Replace the content words by schematic letters; the relation must survive.
conceptual_rule	A meaning postulate of the language: ‚Junggeselle $\rightarrow$ unverheirateter Mann‘, ‚Witwe $\rightarrow$ Frau, deren Ehemann gestorben ist‘, ‚Brief schreiben $\rightarrow$ etwas schreiben‘.	State the rule as ‚ $A \rightarrow B'$ ‘. It must hold by the meaning of the words, independent of the facts.
synonymy	Substitution of a synonym or a historical-orthographic variant with the same	The substituted words must have the same extension in every situation.

Grund	Beschreibung	Prüfung
	sense (vnd/und, Mathematicus/Matematicus).	on; near-synonyms (Astronom/Astrologe) do not qualify.
hyponymy	Taxonomic subordination: Mars → Planet, Brief → Schriftstück. Direction depends on monotonicity of the position.	Name the hypernym chain and check whether the position is upward- or downward-monotone.
antonymy	Complementary antonyms (lebendig/tot) give contradictories; scalar or multi-member antonyms (heiß/kalt, Graz/Prag) give contraries.	Decide whether the opposition has a middle or third value.
converse_relation	Converse relational terms and voice changes: ‚A schreibt an B‘ ↔ ‚B erhält von A einen Brief‘; ‚älter als‘ ↔ ‚jünger als‘.	Check that argument roles are mapped, not merely swapped.
numeric_temporal	Arithmetic and calendar consequences: ‚im März 1598‘ → ‚im Jahr 1598‘ → ‚vor 1600‘; ‚12 Beobachtungen‘ → ‚mehr als zehn Beobachtungen‘.	Compute the consequence; coarsening is entailment, refining is entailed_by, a disjoint value is contrary.
part_whole_place	Mereological and geographic inclusion: ‚in Graz‘ → ‚in der Steiermark‘; ‚im ersten Kapitel‘ → ‚im Buch‘.	The inclusion must be fixed for the time the sentence speaks about; otherwise it is world_knowledge.
world_knowledge	A factual background premise (not a conceptual rule) is needed. Allowed only when the premise is stated in ‚rule‘. Classifiers are instructed to treat such pairs as independent unless told otherwise; the generator should use this basis sparingly and mark it.	Write the premise; if a competent reader could reasonably doubt it, prefer ‚independent‘.
lexical_overlap	Surface similarity without a valid rule (used for near misses and independents).	Shared words and names, no rule.
attitude_context	The content sits under a non-factive attitude or speech verb (glaubte, behauptete, hoffte) or a modal, which blocks consequence to the bare content.	Factive verbs (wusste, bewies) transmit truth; non-factive ones do not.

3.6 Beinahe-Treffer

Typ	ahmt nach	Gold meist	Beschreibung
converse_direction	entails	entailed_by	Specialisation instead of generalisation: hyponym where a hypernym would be needed, added detail.
overlap_no_rule	equivalent	independent	High lexical overlap, but one load-bearing element is replaced so that no rule connects S and P.

Typ	ahmt nach	Gold meist	Beschreibung
false_synonym	equivalent	independent, entailed_by, entails	Near-synonym that differs in extension: Astronom/Astrologe, verlobt/verheiratet, Brief/Bericht.
role_swap	equivalent	independent, contrary	Arguments of a relation exchanged: ‚Kepler schrieb an Tycho‘ vs. ‚Tycho schrieb an Kepler‘.
quantifier_shift	entails	entailed_by, independent, contrary	einige → alle, viele → die meisten, ein → genau ein.
modal_shift	equivalent	entailed_by, independent, entails	Actuality vs. possibility or obligation: ‚tat‘ vs. ‚konnte tun‘ vs. ‚sollte tun‘.
temporal_shift	equivalent	contrary, independent	Date, tense or sequence changed for a singular event.
entity_substitutio	equivalent	independent, contrary	Another person, place or work with a similar role.
negation_scope	contradictio	independent, entails, contrary	A negation that does not negate the whole Aussage: ‚nicht nur‘, narrow-scope negation on a modifier, negated attitude.
contrary_as_contra	contradictio	contrary	An incompatible alternative that leaves a third possibility: another place, another value.
attitude_drop	entails	independent	Dropping a non-factive attitude: ‚X glaubte, dass p‘ vs. ‚p‘. Also adding one.
missing_premise	entails	independent	P would follow only with an unstated factual premise.

Die **Fallenquote** eines Laufs ist der Anteil der Beinahe-Treffer, die der Klassifikator genau mit der Relation belegt, die sie nachahmen.

3.7 Zulassung als Aussage

Zugelassen:

- declarative main clauses, also with subordinate clauses
- reports of attitudes or speech (‚Kepler schrieb, dass ...‘) — the report itself is the Aussage
- historical spelling and Latin/French/English text in the source language

Ausgeschlossen:

- questions, imperatives, wishes, greetings, salutations and closings (‚Lieber Herr ...‘, ‚Vale.‘)
- headings, captions, table rows, bibliographic references, page or line numbers
- fragments cut by a page break (starting mid-sentence or ending without a finite clause)
- performatives without a truth value (‚Hiermit danke ich ...‘) and pure exclamations

4 Aufbau

```

PDAI-Seiten-XML / frischer index.duckdb
├── harvest    Satzteilung · Aussage-Vorfilter · Herkunft (Pfad, OCR-Modell, sha256, mtime)
├── ▼ data/sentences.json    Kandidatensätze
├── │ probe        LLM: Aussage bestätigen · Proben je Zielrelation · Beinahe-Treffer
├── ▼ data/benchmark.json    Sätze mit attributes = Gold je Probe
├── │ bench        Probe × [Quellsatz + Ablenker] → Klassifikator
├── ▼ runs/bench.duckdb      runs · predictions · Sicht leaderboard
runs/<run_id>.json            vollständiger Laufbericht
├── │ audit        Abweichungen auf Quellpaaren → runs/audit.jsonl (Goldprüfung)
├── │ docs         scripts/docs_generate.py → docs/_generiert/ → dieses Dokument

```

Pfad	Inhalt
config/relations.json	Relationskatalog: Labels, Verfahren, Gründe, Beinahe-Typen, Zulassung
prompts/probe_generation.md	Anweisung an den Generator (Katalog wird eingesetzt)
prompts/classify_pair.md	Klassifikator: ein Paar je Aufruf
prompts/classify_list.md	Klassifikator: eine ganze Kandidatenliste je Aufruf
src/famulus_logic/harvest.py	Ernte aus PDAI, Satzteilung, Herkunft, Frischeprüfung
src/famulus_logic/aussage.py	heuristischer Aussage-Vorfilter
src/famulus_logic/probes.py	Probenerzeugung, Prüfung, Cache
src/famulus_logic/bench.py	Aufgabenbau, Lauf
src/famulus_logic/metrics.py	Fehlermatrix und Kennzahlen
src/famulus_logic/store.py	DuckDB-Schema der Läufe
src/famulus_logic/classifiers/	Schnittstelle, lexikalische Basislinie, LLM-Klassifikator
scripts/docs_generate.py	erzeugt die Tabellen dieses Dokuments

5 Ernte der Sätze

5.1 Quellen

Je Dokument wird bevorzugt die Tabelle `sentences` aus `<doc>/index.duckdb` gelesen — aber **nur, wenn sie frisch ist**: ihr `mtime` darf nicht älter sein als irgendeine Seiten-OCR-XML des Dokuments. Ein veralteter Index wird übergangen und als `index_stale` gezählt (PDAI-Integrität: Vorhandensein ist nicht Frische). Sonst wird die Seiten-XML gelesen: `<sammlung>/<doc>/pages/page_NNN/page_NNN.xml`, Text aller `<p>` in `<main>`. Randzeilennummern (`<marginal>`), Fußnotenzeichen, Seitenmarken, Abbildungen und Kopfzeilen werden entfernt; `<unclear>` wird als (?) markiert und führt zur Abweisung. Der Text selbst wird nicht verändert, auch nicht entsilbt — die Transkription bleibt wörtlich.

Mit `--glob` lassen sich nicht-kanonische Ablagen lesen (z. B. Qualitäts-Schnappschüsse mit mehreren OCR-Lesungen derselben Seite); je (Sammlung, Dokument, Seite) wird dann nur eine Variante genommen.

Jeder Satz erhält eine stabile Kennung `<sammlung>/<doc>/p<seite>/s<nr>—<sha1>` und seine Herkunft: Sammlung, Dokument, Seite, gedruckte Seite, Pfad, Textschicht (`ocr` oder `translation`), OCR-Modell, sha256 und mtime der XML. Der vorangehende Satz wird als Kontext mitgeführt, damit der Generator Pronomen auflösen kann.

5.2 Satzteilung und Vorfilter

Geteilt wird an `.`, `!`, `?` vor einem Großbuchstaben oder einer Jahreszahl, außer nach Abkürzungen (vgl., Nr., S., Bd., z. ...), Initialen, römischen Zahlen und Ordinalzahlen (*am 12. März*); eine Zahl nach einer Zitierabkürzung (S. 4.) beendet den Satz.

Der Vorfilter (`heuristic/1`) weist nur ab, was sicher keine Aussage ist: Fragen, Ausrufe, fehlende Schlusszeichen, Kleinschreibung am Anfang (Seitenbruch-Fragment), weniger als 6 oder mehr als 80 Wörter, geringer Buchstabenanteil (Tabellen, Register), Anreden und Formeln, Verweise, Literaturangaben (*Kremer, 2004 ...*). Was durchkommt, ist **Kandidat**; die Entscheidung trifft der Generator.

5.3 Ernte des Piloten

Das Laufwerk golden4T war nicht eingehängt. Der Pilot erntete deshalb aus den lokalen Qualitäts-Schnappschüssen unter `~/ai-mac-air/pdai/collections`: Kirch (Doppellesung, deutsche Sekundärliteratur mit Quellenzitaten), Kepler kgw-13 (Briefe, frühneuhochdeutsch) und Hevelius (Pilot; diese Seiten sind die **englische Übersetzungsschicht**, nicht die OCR).

Erstellt 2026-10-11T11:53:48+00:00 · Wurzel
/Users/friedrichstr/ai-mac-air/pdai/collections · 42 Dokumente · Höchstzahl je Dokument 150 · Seed 13 · Filter `heuristic/1`

Muster:

- `kirch/quality/dual-read/*/page_*.xml`
- `kepler-prim/quality/deepseek-v41-assessment/*/xml-kgw13/sample20/page_*/page_*.xml`
- `hevelius/pilot/*/page_*.xml`

Schritt	Anzahl
Seiten gelesen	3771
Seitenvarianten übersprungen (Doppellesung)	4547
Sätze gesehen	75707
Kandidaten behalten (nach Höchstzahl je Dokument)	2524

Abweisungsgründe des Vorfilters (ein Satz kann mehrere haben):

Grund	Sätze
<code>too_short</code>	28832
<code>low_letter_ratio</code>	18175
<code>no_final_punctuation</code>	16638
<code>bibliography</code>	2940
<code>starts_lowercase_fragment</code>	1491

Grund	Sätze
question	640
too_long	432
exclamation_or_imperative	96
reference	95
salutation_or_formula	14

Kandidaten nach Sammlung, Textschicht und Sprache:

Sammlung	Schicht	Sprache	Sätze
hevelius	translation	en	569
kepler-prim	ocr	orig	249
kirch	ocr	de	1706

6 Erzeugung der Proben

6.1 Verfahren

Je Kandidatensatz ein Aufruf des Generators (Standard `deepseek-v4.1-flash:cloud` über Ollama, Temperatur 0,3). Die Anweisung `prompts/probe_generation.md` enthält den vollständigen Katalog, die Beinahe-Typen und die Zulassungsregeln und verlangt:

- das Urteil, ob S eine Aussage ist, mit Begründung, und eine Paraphrase;
- je eine Probe für jede Zielrelation, wo eine natürliche Probe existiert: `equivalent`, `entails`, `entailed_by`, `contradictory`, `contrary`, `independent`;
- mindestens vier Beinahe-Treffer verschiedener Typen, je mit wahrer Relation, nachgeahmter Relation und Typ;
- für jede Probe Grund, Regel, die vier Zellen und eine Begründung in einem Satz. Proben sind in der Sprache von S geschrieben; moderne Schreibung ist erlaubt.

6.2 Prüfung und Gold-Status

Jede Probe wird gegen den Katalog geprüft: unbekannte Relation, leerer Ausdruck oder unbekannter Beinahe-Typ verwerfen sie. Stimmen die vier Zellen des Generators mit seinem Label überein, erhält sie `gold_status = generated`, sonst `inconsistent_cells` — solche Proben gehen nicht in Läufe ein. Nach menschlicher Prüfung wird der Status `verified` oder `human`.

Antworten werden je (Satz, Modell, Prompt-Hash, Katalog-Hash) zwischen gespeichert (`data/cache/probes/`), ein unterbrochener Lauf setzt fort.

6.3 Benchmark des Piloten

Erzeugt 2026-10-11T12:03:36+00:00 mit `deepseek-v4.1-flash:cloud` (Temperatur 0.3, mindestens 4 Beinahe-Treffer je Satz) · Katalog v1.0.0 `5c74dc5e0a9e` · Prompt `cbcdf5e76607` · Benchmark-sha256 `cec91bf9a879`

Größe	Wert
Sätze vorgelegt	60
als Aussage bestätigt	39
vom Modell abgewiesen	21

Größe	Wert
Proben	467
davon Beinahe-Treffer	233
Proben mit inkonsistenten Zellen	0
ungültige Proben verworfen	0

Gold-Relationen, getrennt nach Zielproben und Beinahe-Treffern:

Gold-Relation	Zielproben	Beinahe-Treffer	Summe
equivalent	39	0	39
entails	39	25	64
entailed_by	39	73	112
contradictory	39	0	39
contrary	39	22	61
subcontrary	0	1	1
independent	39	112	151

Grund	Proben
formal	204
lexical_overlap	144
conceptual_rule	35
synonymy	29
antonymy	18
numeric_temporal	17
hyponymy	12
attitude_context	4
converse_relation	3
part_whole_place	1

Beinahe-Typ	Proben
overlap_no_rule	37
false_synonym	37
converse_direction	34
modal_shift	29
quantifier_shift	28
role_swap	27
negation_scope	15

Beinahe-Typ	Proben
temporal_shift	9
missing_premise	5
attitude_drop	4
contrary_as_contradictory	4
entity_substitution	4

Der Generator hat ein gutes Drittel der Kandidaten als Nicht-Aussagen abgewiesen — vor allem Registereinträge und Listen aus den Kirch-Bänden, die der Vorfilter nicht erkennen kann. Beispiele:

Samm- lung	Satz	Begründung des Modells
hevelius	Add this Diamond to your Crown, and do not weary of adding clarity upon clarity to it.	S is an imperative/request („Add this Diamond... and do not weary...“), not a declarative sentence, so it is not truth-a
hevelius	To the Most Serene and Most Powerful Prince and Lord To Lord Ludovico XIV. Most Christian King of France and N...	S is a dedication/salutation addressed to a prince, not a declarative sentence with a truth value; it lacks a finite cla
kepler-prim	His itaque ulterius ingenij nervos applica; et si omnimodam consonantiam, quae nullâ ex parte claudicet, et in...	The sentence begins with an imperative main clause („applica“) and is an exhortation, not a declarative sentence with a
kirch	Henecka, Hans Peter: Soziale Bedingungen von Festen.	S ist ein bibliographischer Eintrag (Autor und Titel) und keine Aussage; ihm fehlt ein finites Prädikat und ein Wahrheit
kirch	Meise, Helga: Schreibkalender und Autobiographie in der Frühen Neuzeit.	Bibliographische Angabe (Autor: Titel), keine wahrheitsfähige Aussage.
kirch	Tersch, Harald: Der Schreibkalender als Selbstzeugnis.	Bibliographische Angabe (Autor: Titel), kein deklarativer Satz mit Wahrheitswert.
kirch	Gottfried, Johann Ludwig: Neue Archontologia cosmica, Das ist Beschreibung aller Kayserthumben, Königreichen u...	S is a bibliographic reference (author, title, place, year), not a declarative sentence with a truth value.
kirch	Nr. 1 (Berlin, 1924); ders.: Die Meteorologie in den Deutschen Flugschriften und Flugblättern des XVI. Jahrhun...	Der Satz ist ein bibliographischer Verweis bzw. eine Literaturangabe und keine wahrheitsfähige Aussage.

7 Benchmark-Aufgaben

Aus jeder Probe mit zugelassenem Gold-Status wird eine Aufgabe:

- **Quellpaar:** die Probe mit ihrem eigenen Satz, Gold aus dem Benchmark;

- **Ablenker:** `k_random` Sätze (Standard 4) aus **anderen** Dokumenten und `k_same_doc` (Standard 1) aus **demselden** Dokument, Gold independent mit `gold_source = assumed:other_doc` bzw. `assumed:same_doc`.

Die Reihenfolge der Kandidaten wird je Probe zufällig, aber reproduzierbar (Seed) gemischt. `--max-probes` zieht eine Stichprobe über den ganzen Benchmark; `--near-miss only|none` beschränkt auf oder schließt Beinahe-Treffer aus.

⚠ Warnung

Das Ablenker-Gold ist **angenommen**. Ablenker aus demselben Dokument sind schwerer und gelegentlich wirklich verwandt; sie werden in den Kennzahlen getrennt ausgewiesen.

8 Klassifikatoren

Ein Klassifikator erhält eine Probe und eine Kandidatenliste und gibt je Kandidat ein Urteil zurück: Label, optional die vier Zellen, Latenz, Tokens, Fehler. Paar-Klassifikatoren werden über die Liste iteriert.

lexical (Version 1)

Basislinie aus Wortüberlappung (Jaccard ohne Funktionswörter), Enthaltensein und Parität der Negationswörter (deutsch, lateinisch, französisch, englisch). Sie markiert die Untergrenze: was Oberflächen ähnlichkeit allein leistet.

llm:<modell> (Version 1)

Beliebiges OpenAI-kompatibles Endpunkt-Modell (Standard Ollama `http://localhost:11434/v1`). Die Anweisung enthält den gerenderten Katalog mit Beispielen. `mode=list`: ein Aufruf je Aufgabe mit nummerierter Kandidatenliste — das ist die eigentliche Fragestellung „steht P zu irgendeinem dieser Sätze in Beziehung?“. `mode=pair`: ein Aufruf je Paar. `decide=cells` leitet das Label aus den Zellen ab, `decide=label` nimmt das genannte; beide Werte werden je Urteil aufbewahrt.

Als Eigenschaften eines Laufs werden festgehalten: Modell, Endpunkt, Temperatur, Seed, JSON-Modus, Modus, Entscheidungsart, Beispiele an/aus, Name und sha256 der Prompt-Vorlage, Version und sha256 des Katalogs.

9 Kennzahlen

Kennzahl	Definition
Fehlermatrix (fein)	7×7 , Zeilen Gold, Spalten Vorhersage, über alle Paare (Quellpaare und Ablenker)
Fehlermatrix (grob)	3×3 über die Familien <code>follows</code> / <code>opposes</code> / <code>none</code>
Genauigkeit	Anteil richtiger Paare — durch die vielen Ablenker geschönt
Makro-F1	Mittel der F1-Werte über die im Gold vorkommenden Labels; Hauptkennzahl
gewichtetes F1	F1 gewichtet mit Gold-Häufigkeit

Kennzahl	Definition
Cohens κ	Übereinstimmung mit dem Gold, bereinigt um Zufall
verwandt-F1	binär: irgendeine Relation gegen <code>independent</code>
Quellpaare	Genauigkeit nur auf den Paaren Probe $\leftrightarrow$ eigener Satz
Zielproben / Beinahe-Treffer	Genauigkeit auf den Quellpaaren, getrennt
Fallenquote	Anteil der Beinahe-Treffer, die als ihre nachgeahmte Relation klassifiziert werden
Ablenker-Fehlalarm	Anteil der Ablenker, denen eine Relation zugesprochen wird
Listen-Erkennung F1	je Aufgabe: Ist irgendein Kandidat verwandt? Gold gegen Vorhersage
Lokalisierung	unter Aufgaben mit verwandtem Kandidaten: Quellpaar richtig klassifiziert
exakte Liste	ebenso, und alle Ablenker richtig
nach Grund / Beinahe-Typ	Genauigkeit und Fallenquote je <code>basis</code> und <code>near_miss_type</code>
Leistung	Wandzeit, Paare/s, Latenz je Paar (Mittel, p50, p95), Tokens ein/aus, Fehlerzahlen

Im Listenmodus wird die Latenz eines Aufrufs gleichmäßig auf seine Paare verteilt; `call_latency_s` bleibt je Urteil erhalten.

10 Die Lauftabelle

`runs/bench.duckdb` enthält:

- `runs` — eine Zeile je Lauf: Kennung, Zeit, Notiz, Klassifikator, Version, Modell, Modus, Entscheidungsart, alle Eigenschaften (JSON), Benchmark-Pfad und `-sha256`, Katalog-Version und `-sha256`, Aufgabenparameter, Anzahl Aufgaben und Paare, alle Kennzahlen als Spalten, feine und grobe Fehlermatrix (JSON), vollständige Kennzahlen (JSON), Umgebung (Rechner, Python, Git-Commit und ob der Arbeitsbaum geändert war), `self_graded`;
- `predictions` — eine Zeile je Paar und Lauf: Gold, Vorhersage, Gold-Quelle, Beinahe-Typ, nachgeahmte Relation, Grund, Latenz, Fehler, Zellen;
- Sicht `leaderboard` — nach Makro-F1 sortiert.

`self_graded` ist wahr, wenn das klassifizierende Modell zugleich das Gold erzeugt hat. Solche Läufe messen Selbstkonsistenz, nicht Richtigkeit.

11 Ergebnisse des Piloten

11.1 Läufe

Lauf	Notiz	Zeit
20261011T120341Z-lexical-baseline-591f20ff	pilot floor	2026-10-11 14:03

Lauf	Notiz	Zeit
20261011T120353Z- llm-f0ad2158	pilot: kimi-k3 list/cells, 60 probes	2026-10-11 14:03
20261011T121601Z- llm-f0ad2158	review: kimi-k3 list/cells, 60 sampled probes	2026-10-11 14:16
20261011T121601Z- llm-5a18b15b	review: deepseek-v4.1-flash list/cells, 60 sampled probes (generator model)	2026-10-11 14:16

11.2 Rangliste

Zeit	Klassifikator	selbst	Proben	Paare	Makro-F1	κ	grob	Quellpaare	Fallen	Lokal.	Wands	Paare/s
14:16	kimi-k3	nein	60	338	0.823	0.820	0.967	0.767	0.036	0.805	545.7	0.62
14:03	kimi-k3	—	60	323	0.789	0.789	0.969	0.733	0.233	0.769	640.9	0.50
14:16	deepseek-v4.1-flash	ja	60	338	0.760	0.757	0.962	0.717	0.071	0.634	1076.3	0.31
14:03	lexical-baseline	—	467	2607	0.297	0.338	0.904	0.336	0.270	0.165	0.0	54438

Spalten: Zeit = Startzeit (volle Kennung in der Tabelle darüber); selbst = `self_graded`; grob = Genauigkeit der groben Matrix; Fallen = Fallenquote; Lokal. = Lokalisierung; Wands = Wandzeit.

11.2.1 Auswertung

Die beiden aktuellen LLM-Läufe (kimi-k3 und deepseek-v4.1-flash) liefern auf **derselben Stichprobe**: 60 Proben aus 33 Sätzen, je ein Quellpaar und fünf Ablenker, zusammen 338 Paare, Listenmodus, Label aus den Zellen abgeleitet. Der erste kimi-Lauf (...120353Z...) stammt aus einer älteren, nicht gestreuten Stichprobe (60 Proben aus wenigen Sätzen) und ist nur zur Dokumentation aufgeführt.

Ergebnis. kimi-k3 erreicht Makro-F1 0,823 (κ 0,820), deepseek-v4.1-flash 0,760 (κ 0,757), die lexikalische Basislinie 0,297. Der Abstand der LLMs zur Basislinie ist groß und eindeutig; der Abstand zwischen den beiden LLMs ist es bei 59 bewerteten Quellpaaren nicht.

Selbstbewertung. deepseek-v4.1-flash hat das Gold erzeugt. Als Klassifikator stimmt es nur auf 72 % der Quellpaare mit seinem eigenen Gold überein — weniger als kimi (77 %). Das Modell reproduziert seine Labels also nicht, wenn es dieselbe Frage im Listenmodus ohne die Bauabsicht gestellt bekommt. Das spricht gegen eine bloße Selbstbestätigung, aber auch dafür, dass ein Teil des Goldes instabil ist.

Übereinstimmung der Modelle. Untereinander stimmen kimi und deepseek auf 80 % der Quellpaare überein (κ 0,83 über alle gemeinsamen Paare) — mehr als jedes der beiden mit dem Gold. Auf 8 von 59 Quellpaaren sind sich beide einig **gegen** das Gold. Diese Fälle sind keine sicheren Goldfehler: Bei *Non solum tuas litteras expectavi* (Gold `entailed_by`, beide Modelle `independent`) ist das Gold vertretbar, denn „nicht nur deine Briefe habe ich erwartet“ schließt ein, dass die Briefe erwartet wurden. Beim Thomasius-Satz (Gold `equivalent`, beide `entails`) lässt die Probe einen Nebensatz weg und ist eher schwächer — hier spricht mehr für die Modelle. Die Liste ist deshalb die **Prüfliste**, nicht die Korrektur.

Wo die Fehler liegen. In beiden Läufen ist `entailed_by` (P stärker als S) die schwächste Klasse: kimi verfehlt 5 von 16, deepseek 9 von 16, meist zugunsten von `independent`. Das sind überwiegend Beinahe-Treffer vom Typ `converse_direction` und `quantifier_shift`, bei denen die Probe mehr behauptet als der Satz — beide Modelle lesen „mehr“ oft als „anderes“. `contradictory` und `contrary` werden von kimi fehlerfrei erkannt; deepseek verwechselt sie einmal untereinander. Fehlalarme auf Ablenkern gibt es in beiden Läufen nicht.

Beinahe-Treffer. Die Fallenquote ist niedrig (kimi 0,036, deepseek 0,071): Die Modelle fallen selten auf die *nachgeahmte* Relation herein. Sie verfehlen Beinahe-Treffer häufiger, indem sie sie für unabhängig halten, wo das Gold eine Relation in der Gegenrichtung angibt. Beim alten kimi-Lauf lag die Fallenquote bei 0,233 — die Kennzahl hängt bei dieser Stichprobengröße stark an der Auswahl.

Leistung. kimi brauchte 546 s Wandzeit (0,62 Paare/s, Median 5,8 s je Paar bei sechs Kandidaten je Aufruf). deepseek antwortet je Aufruf schneller (Median 3,1 s je Paar), seine Wandzeit von 1076 s ist aber von **einer** Aufgabe bestimmt, deren Aufruf dreimal nach 300 s an einem 502 von Ollama Cloud scheiterte. Diese 6 Paare zählen im Lauf als `call_failed` und als `independent` vorhergesagt; im Modellvergleich sind sie ausgeschlossen. Die Basislinie ist um Größenordnungen schneller, aber unbrauchbar.

Schluss für den Piloten. Der Prüfstand trennt Klassifikatoren klar von der Oberflächenähnlichkeit und zeigt, wo sie scheitern (Richtung der Folgerung bei stärkeren Proben). Für eine Rangfolge zwischen starken LLMs reicht die Stichprobe nicht; vorher sollte das Gold auf den uneinigen und den einstimmig abweichenden Quellpaaren geprüft werden.

11.3 Fehlermatrizen der aktuellen LLM-Läufe

kimi-k3:cloud · 20261011T121601Z-llm-f0ad2158 · review: kimi-k3 list/cells, 60 sampled probes

Eigenschaften: `model=kimi-k3:cloud`, `base_url=http://localhost:11434/v1`, `temperature=0.0`, `seed=7`, `json_mode=True`, `max_tokens=None`, `mode=list`, `decide=cells`, `examples=True`, `prompt_template=classify_list.md`, `relations_version=1.0.0`

Gold \ vorhergesagt	equivalent	entails	entailed by	contradictory	contrary	subcontrary	independent
equivalent	2	1	0	0	0	0	0
entails	0	6	0	0	0	1	1
entailed by	1	1	11	0	0	0	3
contradictory	0	0	0	7	0	0	0
contrary	0	0	0	0	7	0	0
subcontrary	0	0	0	0	0	0	0
independent	0	4	0	0	2	0	291

Label	Präzision	Recall	F1	Gold	vorhergesagt
equivalent	0.667	0.667	0.667	3	3
entails	0.500	0.750	0.600	8	12
entailed_by	1.000	0.688	0.815	16	11
contradictory	1.000	1.000	1.000	7	7
contrary	0.778	1.000	0.875	7	9
subcontrary	0.000	—	—	0	1
independent	0.986	0.980	0.983	297	295

Grobe Matrix (Familien):

Gold \ vorhergesagt	follows	opposes	none
follows	22	1	4
opposes	0	14	0
none	4	2	291

Leistung	Wert
Wandzeit s	545.7
Paare/s	0.619
Latenz Mittel / p50 / p95 s (je Paar)	9.172 / 5.826 / 25.465
Tokens ein / aus	179596.0 / 256006.0
Fehler	keine

deepseek-v4.1-flash:cloud · 20261011T121601Z-llm-5a18b15b · review: deepseek-v4.1-flash list/cells, 60 sampled probes (generator model)

Eigenschaften: model=deepseek-v4.1-flash:cloud, base_url=http://localhost:11434/v1, temperature=0.0, seed=7, json_mode=True, max_tokens=None, mode=list, decide=cells, examples=True, prompt_template=classify_list.md, relations_version=1.0.0

Gold \ vorhergesagt	equivalent	entails	entailed by	contradictory	contrary	subcontrary	independent
equivalent	2	1	0	0	0	0	0
entails	0	6	0	0	0	1	1
entailed by	0	2	7	0	0	0	7
contradictory	0	0	0	7	0	0	0
contrary	0	0	1	1	4	0	1

Gold \ vorhergesagt	equivalent	entails	entailed by	contradictory	contrary	subcontrary	independent
subcontrary	0	0	0	0	0	0	0
independent	1	1	0	0	0	0	295

Label	Präzision	Recall	F1	Gold	vorhergesagt
equivalent	0.667	0.667	0.667	3	3
entails	0.600	0.750	0.667	8	10
entailed_by	0.875	0.438	0.583	16	8
contradictory	0.875	1.000	0.933	7	8
contrary	1.000	0.571	0.727	7	4
subcontrary	0.000	—	—	0	1
independent	0.970	0.993	0.982	297	304

Grobe Matrix (Familien):

Gold \ vorhergesagt	follows	opposes	none
follows	18	1	8
opposes	1	12	1
none	2	0	295

Leistung	Wert
Wandzeit s	1076.3
Paare/s	0.314
Latenz Mittel / p50 / p95 s (je Paar)	5.276 / 3.128 / 8.534
Tokens ein / aus	164607.0 / 232912.0
Fehler	call_failed: 6

11.4 Lexikalische Basislinie

lexical-baseline · 20261011T120341Z-lexical-baseline-591f20ff · pilot floor

Eigenschaften: equiv=0.75, overlap=0.5, contain=0.9

Gold \ vorhergesagt	equivalent	entails	entailed by	contradictory	contrary	subcontrary	independent
equivalent	4	3	0	0	0	0	32
entails	12	18	1	0	0	0	33

Gold \ vorhergesagt	equivalent	entails	entailed by	contradictory	contrary	subcontrary	independent
entailed by	33	2	5	8	0	0	64
contradictory	0	0	0	25	0	0	14
contrary	4	6	0	9	0	0	42
subcontrary	0	0	0	1	0	0	0
independent	41	4	1	0	0	0	2245

Label	Präzision	Recall	F1	Gold	vorhergesagt
equivalent	0.043	0.103	0.060	39	94
entails	0.545	0.281	0.371	64	33
entailed_by	0.714	0.045	0.084	112	7
contradictory	0.581	0.641	0.610	39	43
contrary	—	0.000	—	61	0
subcontrary	—	0.000	—	1	0
independent	0.924	0.980	0.951	2291	2430

Grobe Matrix (Familien):

Gold \ vorhergesagt	follows	opposes	none
follows	78	8	129
opposes	10	35	56
none	46	0	2245

Leistung	Wert
Wandzeit s	0.0
Paare/s	54437.632
Latenz Mittel / p50 / p95 s (je Paar)	0.000 / 0.000 / 0.000
Tokens ein / aus	— / —
Fehler	keine

11.5 Beinahe-Treffer nach Typ

Beinahe-Typ	kimik3:cloud (n · Treffer · Falle)	deepseek-v4.1-flash:cloud (n · Treffer · Falle)
contrary_as_contradictory	1 · 1.000 · 0.000	1 · 1.000 · 0.000

Beinahe-Typ	kimi-k3:cloud (n · Treffer · Falle)	deepseek-v4.1-flash:cloud (n · Treffer · Falle)
converse_direction	4 · 0.500 · 0.000	4 · 0.500 · 0.000
entity_substitution	2 · 1.000 · 0.000	2 · 1.000 · 0.000
false_synonym	3 · 0.333 · 0.000	3 · 0.333 · 0.333
modal_shift	1 · 1.000 · 0.000	1 · 1.000 · 0.000
negation_scope	3 · 0.667 · 0.000	3 · 0.333 · 0.000
overlap_no_rule	5 · 0.800 · 0.000	5 · 1.000 · 0.000
quantifier_shift	5 · 0.600 · 0.200	5 · 0.400 · 0.200
role_swap	4 · 0.500 · 0.000	4 · 1.000 · 0.000

11.6 Genauigkeit nach Grund

Grund	kimi-k3:cloud (n · Treffer)	deepseek-v4.1-flash:cloud (n · Treffer)
antonymy	2 · 1.000	2 · 0.500
conceptual_rule	3 · 1.000	3 · 0.667
converse_relation	1 · 1.000	1 · 1.000
formal	32 · 0.750	32 · 0.656
lexical_overlap	18 · 0.722	18 · 0.889
numeric_temporal	2 · 1.000	2 · 0.500
synonymy	2 · 0.500	2 · 0.500

11.7 Vergleich der Modelle

Auf den gemeinsamen Paaren der aktuellen LLM-Läufe (gleiche Stichprobe, gleicher Seed):

6 Paare mit fehlgeschlagenem Aufruf (in irgendeinem Lauf) sind ausgeschlossen.

Paar	gemeinsame Paare	Übereinstim- mung	Quellpaare	Übereinst. Quellpaare	κ
kimi-k3:cloud ↔ deep- seek-v4.1- flash:cloud	332	0.964	59	0.797	0.830

Quellpaare (Probe ↔ eigener Satz)	Anzahl
alle Modelle treffen das Gold	39
Modelle uneinig	12
alle Modelle einig gegen das Gold (Gold verdächtig)	8
gemeinsam bewertet	59

Wo alle Modelle übereinstimmen und dem Gold widersprechen, ist das Gold verdächtig — diese Fälle gehören zuerst in die Prüfung:

S	P	Gold	alle Modelle	Beinahe-Typ
On the ground of the observation of phenomena alone and of scientific hypotheses, a dialogue on the nature of comets, their forms and their ...	A dialogue on the nature of comets was possible and very desirable on the basis of observation of phenomena and scientific hypotheses, and H...	entailed_independent	converse_direction	
Maximo cum desiderio, Domine frater amiciss: tuas expectaui literas.	Non solum tuas litteras expectavi.	entailed_independent	negation_scope	
Quod tandem aliquando Elephas meus peperit; nec interea temporis ulla libello meo difficultas a scripturae sacrae defensoribus, quod metueba...	Nec interea temporis ullum problema libello meo a scripturae sacrae defensoribus obiectum fuit.	independent	entails	false_synonym
Cum ergò EGB ex observatione altitudinis Solaris sit 36°. 20' (·nam paulo post finem fuit 36°. 10'·) et BIG rectus, ergo IBG erit residuus, ...	EGB est 36°20', paulo post finem EGB fuit 36°10', BIG est rectus, IBG est 53°40', et EGB est 36°20' et Sol in Geminis est.	entailed_independent	converse_direction	
Porro ut aliquid etiam pro te afferam: viderj non injuria possit illa summa 480 annorum obiter et sine curâ collecta.	Illa summa 480 annorum collecta est.	entails_independent	–	
Das wichtigste Anliegen des herausragenden Frühaufklärers Christian Thomasius, die Bekämpfung der Vorurteile, was die Kritik am Aberglauben ...	Das wichtigste Anliegen von Christian Thomasius, die Bekämpfung der Vorurteile einschließlich der Kritik am Aberglauben, lässt sich auch als...	equivalent	entails	–
Das wichtigste Anliegen des herausragenden Frühaufklärers Christian Thomasius, die Bekämpfung der Vorurteile, was die Kritik am Aberglauben ...	Das wichtigste Anliegen von Christian Thomasius, die Bekämpfung der Vorurteile einschließlich der Kritik am Aberglauben, ist bei allen genau...	entailed	entails	–
Hat Nebussaradan das Jüdische Land bekriegt: so ist dagegen zwischen den Sabinern und Tarquinium	Nebussaradan bekriegt das Jüdische Land.	entails	subcontrary	–

S	P	Gold	alle Modelle	Beinahe-Typ
ein guter Friede getroffen/ massen aus dem...				

Uneinigkeit der Modelle auf Quellpaaren, nach Gold:

Gold	kimik3:cloud / deepseek-v4.1-flash:cloud	Anzahl
contrary	contrary / independent	1
contrary	contrary / entailed_by	1
contrary	contrary / contradictory	1
entailed_by	entailed_by / independent	3
entailed_by	equivalent / entails	1
independent	entails / independent	2
independent	contrary / independent	2
independent	entails / equivalent	1

12 Prüfung des Goldes

Das Gold ist erzeugt, nicht geprüft. Der Prüfkreislauf:

1. `famulus-logic audit --run <run_id>` schreibt alle Quellpaare, auf denen ein Lauf vom Gold abweicht, nach `runs/audit.jsonl` — mit S, P, Gold, Vorhersage, Beinahe-Typ, Regel und Begründung des Generators und einem leeren Feld `verdict`.
2. Zuerst die Fälle prüfen, in denen **alle** Modelle gegen das Gold stehen (Abschnitt *Vergleich der Modelle*).
3. Im Benchmark `relation` korrigieren und `gold_status` auf `human` setzen (oder `verified`, wenn ein zweites Modell und ein Mensch zustimmen).
4. Neu laufen lassen; der Benchmark-sha256 im Lauf zeigt, auf welchem Goldstand gemessen wurde.

Beobachtungen aus dem Piloten: Mehrere Abweichungen sind Fehler des Goldes, nicht des Klassifikators — etwa *seal of secrecy* gegen *seal of silence*, vom Generator als `independent` (falsches Synonym) gesetzt, obwohl beide Wendungen dasselbe sagen. Beinahe-Treffer sind die anfälligste Stelle: Der Generator soll eine Falle bauen und behauptet dabei manchmal einen Unterschied, den die Wörter nicht tragen.

13 Bedienung

```
cd ~/ai-mac-air/development/famulus-logic uv sync --extra dev
uv run famulus-logic relations # Katalog anzeigen und prüfen
# Ernte (Standardwurzel $PDAI_ROOT oder /Volumes/golden4T/pdai-collections)
uv run famulus-logic harvest --collections kepler-prim karman-nachlass
--per-doc 200 # Schnappschüsse, solange das Laufwerk fehlt
uv run famulus-logic harvest --root ~/ai-mac-air/pdai/collections \
--glob 'kirch/quality/dual-read/*/page_*.xml' 'hevelius/pilot/*/page_*.xml'
--per-doc 150 # Proben
uv run famulus-logic probe --model deepseek-v4.1-flash:cloud --limit 500
```

```

--workers 8 # Läufe uv run famulus-logic bench --classifier lexical uv
run famulus-logic bench --classifier llm:kimi-k3:cloud --mode list --decide
cells uv run famulus-logic bench --classifier llm:deepseek-v4.1-flash:cloud
--max-probes 60 uv run famulus-logic report # Rangliste
uv run famulus-logic report --run < run_id > # Fehlermatrix eines Laufs
uv run famulus-logic audit
--run < run_id > # Abweichungen zur Goldprüfung # Dieses Dokument uv
run python scripts/docs_generate.py
cd docs && quarto render famulus-logic.qmd --to lettreai-html QUARTO_TYPST
= $( which typst ) quarto render famulus-logic.qmd --to lettreai-typst uv
run --extra dev python -m pytest # Tests

```

Umgebungsvariable	Bedeutung
PDAI_ROOT / PDAI_COLLECTIONS	Wurzel der PDAI-Sammlungen
FAMULUS_LOGIC_LLM_BASE	OpenAI-kompatibler Endpunkt (Standard Ollama <code>http://localhost:11434/v1</code>)
FAMULUS_LOGIC_LLM_KEY	Schlüssel für den Endpunkt (Standard <code>ollama</code>)

14 Grenzen und nächste Schritte

- **Stichprobe.** Die LLM-Läufe des Piloten umfassen 60 Proben aus einem Benchmark von 39 Aussagen; Unterschiede von wenigen Hundertsteln sind nicht belastbar. Für Vergleiche: mehrere hundert Aussagen aus den eigentlichen Sammlungen auf golden4T.
- **Herkunft des Piloten.** Schnappschüsse statt Produktions-OCR; Hevelius in englischer Übersetzung. Für die eigentliche Ernte die kanonischen `pages/`-Ablagen und, wo frisch, `index.duckdb` verwenden.
- **Selbstbewertung.** Der Generator `deepseek-v4.1-flash:cloud` ist auch als Klassifikator gelaufen; sein Lauf ist `self_graded` und misst Konsistenz mit sich selbst. Ein zweiter Generator (z. B. `kimi` oder `glm`) für einen Teil des Benchmarks würde diese Abhängigkeit brechen.
- **Verteilung.** `subcontrary` kommt im Korpus kaum vor; `formal` und `lexical_overlap` dominieren die Gründe. Für Begriffsregeln, Synonymie und Ortseinschluss lohnt eine gezielte Erzeugung (`--targets`, angepasste Anweisung).
- **Ablenker.** `Gold independent` ist angenommen; Fehlalarme auf Ablenkern aus demselben Dokument sollten stichprobenartig geprüft werden.
- **Sprache.** Der Vorfilter ist für Deutsch gebaut; lateinische und französische Seiten brauchen eigene Anrede- und Abkürzungslisten.