

Technical report — Gegenprobe: one schema and a converter for Swiss parliamentary answers

Track 2A — OpenParlData · Online · Team 444 — Roman (+ Claude Code agents, §7) · Evidence site
<https://gegenprobe.million.io>

1. Summary

Question: can one schema hold the business of Switzerland's communal, cantonal and federal parliaments, in German, French and Italian?

Answer: **Yes, if each field says which role its value plays** (e.g. "answer date" = the government's decision, its letter or the submission date), for the 17-field skeleton of the five intervention types in OpenParlData (OPD; 198,594 of its 325,691 affairs); outside interventions the skeleton does not carry over as is (exploratory, 15 affairs).

Delivered: repairs for 9,580 answers keyed on OPD ids, every one from model-free rules; schema v0.2 with the roles; a converter whose records give every value its page and span or its source record, the cantonal and communal ones written by the operator command itself; a nightly job.

Apertus: opt-in; it reads what no rules cover, only picks values printed in the document, code locates them, and a reviewer confirms each; the default runs no model. In 15 parliaments nobody wrote rules for it finds 24/29 answer dates, the rules 2/29 (registered); in 14 it raises author-role recall by 23.4 points over the rules, with precision not worse (registered); 120/120 of its sampled review-queue promotions are the government's answer (registered). Its margin ranks the open documents for review (§3). We use the 70B: on the 52 new-layout documents both read, the local 8B finds 15/21 answer dates (70B 16/21) but dates 11 of 31 answerless documents (70B 4).

Cost: the rules take about 0.5 CPU-hours for all 902,754 documents in OPD's export, about 5 at ten times OPD's volume; review mode costs USD 0.59 per 1,000 residue documents on the 70B (§3).

Terms. **registered:** frozen by hash before the labels; **post hoc:** analysed after them; **role:** which act or actor a value records; **residue:** documents the rules leave open; test codes and exploratory cells: legend below.

maturity	components	
ready to apply · ready to deploy	census and Bern repair lists (relinks: one OPD decision, its hash rule) · nightly job (three real nights on standard GitHub runners + a real two-night Docker pair after a ledger fix: night 2 kept night 1's verdicts)	
audited · tested prototype	source-adapter lists · rules-only converter	
feasibility study · drafts for the eCH group	review mode, whole and virtual affairs · schema v0.1 and v0.2 (P8's roles)	

verdict	fields	example
needs roles	10 (8 robust + 2 fragile; 6 from the role test P8, registered, 4 from OPD's labels, post hoc)	"answer date" is the decision, the letter or, in 7 parliaments, mostly the submission date
body-specific	1 field and 2 constructs	submission date printed in neither document in Vernier (5/5 affairs) and Lausanne (3/5); oral answers; minutes-only councils
untested	6, defined once in v0.1	decision and dispatch dates, Q/A pairing, related affairs, usable text, answer-link status
outside interventions	7 fit, 4 do not apply, 6 need new roles (P11, exploratory, 15 affairs)	the executive as proposer: adopting, endorsing and signing its own proposal

In the converter's 120 records every value carries its role and source, a page and span (1,580 values) or a source record (569 OPD records; 9 federal answering bodies from a one-row committed table), and every gap a typed reason. The answer date's role is in the records: 25 decision and 17 letter or dispatch dates in 28 of the 120 (the blind gold role agrees 32/32); 1,573 of the 1,580 spans also have their box on the PDF page (span_bboxes.jsonl) for a side-by-side view. **What would have meant No** (proposed in review): a field body-specific in 2 or more parliaments for reasons of practice rather than OPD's data, or values outside every named role; the one body-specific field is a split between OPD's date and the documents.

92.1%	of interventions with a recorded government answer would have answer text once the proposed repairs are applied, up from 81.6% today: +10.5 points, all from the 9,580 repairs (the 8,949 row-checked ones reach 91.5%; the 592 agenda joins and 39 Bern rows in review add the rest)
9,580	answers with a model-free repair proposed to OPD, keyed on OPD ids: 3,621 apply now, 2,459 Bern relinks follow OPD's hash decision, 3,500 are fetched; the Bern plan ran on PostgreSQL: 2,975 rows in one transaction that rolls back on any deviation
186	v0.2 records from 26 parliaments validate: the converter's 120 (105 cantonal and communal + 15 federal, 23 parliaments) and P8's 66 gold records (11 parliaments); against blind labels on 36 held-out affairs (P3b) the converter is right on the type 35/36, the submission date 30/34, the submitter 30/36 and the answer date 20/35, 31/35 with Apertus

Test codes. **P1–P8, P10:** the ten pre-registered tests of OPD records and documents (P3b replicates P3), each frozen by hash before its labels except P1's hypothesis script and rules (D4, D12); **P11**, the **federal cell**: locked exploratory cells, without confirmatory hypotheses; **D-numbers**: logged deviations.

Schema. P8 finds roles in authorship, co-signatory lists, the instrument behind a harmonised type and the language. The schema expresses the roles; the converter fills them where the documents print them: 42 of its 142 answer dates carry a split role, and on 5 parliaments its rules never saw the language is right on 60/60 documents and author (name, role) pairs reach precision 39/49 at recall 39/137 (exploratory). Tables, annexes, images, scans and signatures are marked by page range (66 of 105 converter records carry one); in a blind check of 30 documents (108 pages; exploratory, in-sample for the detector) tables are found at page precision 12/12 and recall 12/15, annexes at 6/6 and 6/6, scanned pages at 23/24 and 23/23. Crosswalk to Popolo, Akoma Ntoso, Parla-CLARIN and OParl: docs/mapping.md.

2. Architecture

Census (census/) streams the OPD exports once into a per-affair *answer-link status* and the feasibility matrix, and places every registered gap row (Fig. 1). **Repair** (repair/): deterministic routes per cause, each row checked, as lists keyed on OPD ids; a nightly job re-runs the Bern route. **Converter** (convert/, qa/): OPD metadata and events, then frozen DE/FR/IT anchor rules, then the Apertus closed-set pointer on the residue; modes *rules_only* (default, no model) and *review* (suggestions with page, span and margin, flagged *needs_review*). The operator command `gegenprobe affair` writes the published v0.2 record of a cantonal, communal or federal affair (on the three samples 268/268, 380/380 and 325/325 comparable fields equal). **Schema and validator** (schema/, validate/): Pydantic schema v0.1, normative; a value carries its role, page, span, quote hash, route and rule, a null its typed reason (data/schema/null_reasons.json); QAItem pairs question *n* with its answer passage.

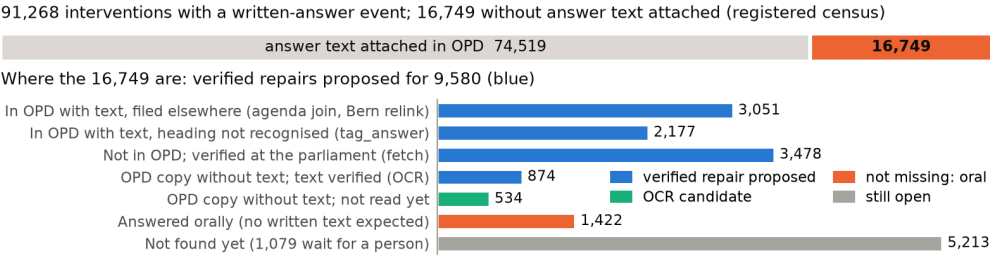


Figure 1. The registered 16,749 placed once each (results/census/corrected.json). Still to find: 5,213 (1,079 of them already waiting for a person).

3. Use of Apertus

Apertus result	test	scope
24/29 answer dates in 15 parliaments nobody wrote rules for; the rules, written for other layouts, 2/29	the date test, P2	registered
+23.4 points author-role recall over the rules in 14 parliaments without rules (50/141 pairs vs 17/141; Holm p = 0.032); precision not worse	the author test, P10 (H1, H2)	registered
120/120 sampled review-queue promotions are the government's answer (Wilson 95% [96.9, 100])	the verifier test, P7 (H1)	registered
12/12 Italian interventions exact on the answer date (10 printed dates, 2 right nulls)	P11	exploratory

Closed form only. Apertus never writes a value. (i) *Verifier* (P7, repair/adapters/adjudicate.py): for an adapter review row it answers "does this document print the executive's answer to this intervention (number, title)?" (yes vs no scored after a JSON prefix, $\theta_A = 8$) and picks the answer date among the printed dates or none ($\theta_B = 10$); a row is promoted only if both pass and the date passes the adapters' own date rule. (ii) *Pointer* (convert/pointer.py, prompts src/prompts/{de, fr, it}): the document and the question, the answer prefilled with a JSON prefix, null and every distinct printed date scored by summed log-probability; null if $\log(\text{null}) - \log(\text{sumexp}(\log(\text{dates}))) > \theta$. Code locates every span (pointer.locate). θ does not transfer across builds, so it is fitted per build on dev documents and frozen ($\theta = 10.40$, AUROC 0.99). **Why closed:** provenance by construction and a ranked review queue (post hoc). Every closed value is a printed date with page and span (62/62, 32/32, 35/35 on P1, P2, P6), while a free prompt's quotes check out on 30/60 (P1) and 17/32 (P2) values; the closed margin ranks the review queue: AUROC for a wrong value 0.73, 0.90, 0.71 on P1, P2, P6, against 0.63, 0.58, 0.53 for the free prompt's quote check, and on the new layouts of P2 above the no-model ranking signals we tested (at most 0.58; the interval of the difference excludes zero). Asked freely without a null option, Apertus dates 15/15 (8B) and 14/15 (70B) Bern interventions that have no answer.

Which model. Evaluation used Apertus v1.5 70B weights served by Public AI under the name `swiss-ai/apertus-v1.5-8b` (inferred from the served fingerprint `...tp8-pp2-dd237840` and 760/792 greedy outputs

identical byte for byte to those of the `-70b` name); every committed gateway response (17,720) comes from that one build, its `model` checked on every call; when Public AI switched the route to another build and back on 12 Oct 2026, the build guard stopped the run, so no result mixes builds (`docs/evaluation.md`).

model	runs	used for	measured
none: <code>rules_only</code> (default)	the operator's machine	every repair; the rules	P1: rules alone 135/154
Apertus 70B ($\theta = 10.40$)	Public AI, or self-hosted	dates, verifier, role reader	P2 24/29; P7 120/120; P10 50/141
Apertus 8B, 4-bit GGUF ($\theta = 16.73$; exploratory)	locally on CPU, no text leaves	date suggestions only	P2 13 vs 0 over the rules; verifier promotes none; roles F1 0.27

Recommendation: `rules_only` unattended; local 8B: date suggestions only, where no text may leave the machine; the review-queue verifier and the role reader need the 70B, for accuracy on infrastructure the operator chooses (`bf16` weights about 140 GB; `docs/runbook.md`).

4. Data and operations

Data. OPD bulk exports (snapshot 6 Oct 2026; CC BY 4.0, "Source: OpenParlData.ch"), Bern open data, Fedlex, parliament sites, fetched politely (≥ 1.2 s per host and process). **Sovereignty and data protection.** No third-party processor by default; a review-model endpoint (Public AI, or the local 8B, where no text leaves the machine) and an alert relay only if the operator sets them. The operator sheet's `nDSG` row names the personal data (office holders named in public documents), the route that discloses text abroad (Public AI: by its published terms a US nonprofit offering no data processing agreement and no storage outside the US) and the cache retention (default 30 days); for public bodies it recommends rules only or the local 8B. A body that may not disclose text abroad goes without only the review-queue verifier and the role reader; a Swiss-hosted 70B is an estimated 1.64 GPU-hours per 1,000 residue documents (`docs/runbook.md`). **Supply chain** (`docs/licences.md`): 8 Python packages in the image under permissive licences, every lock hash-pinned, poppler (GPL) run as a separate program; CycloneDX SBOMs; CI scans every lock for known vulnerabilities (0 failing; 1 accepted with its reason, in the optional local 8B's lock) and the built image (grype), failing on a fixable HIGH or CRITICAL finding: the digest-pinned CPython 3.11.17 base has 0 the image keeps, and the build applies Debian's updates. **Operations.** `make run` and the nightly unit run the image read-only, as a non-root uid, with `--cap-drop ALL`; a network night refuses to start without a contact URL in `GP_USER_AGENT`. OCR fails loudly: without the pinned `deu/fra/ita` models, or on a tesseract error, a scan is a typed `ocr_failed`, never an empty page; with the models all 154/154 P1 documents re-derive from the source PDFs, the 22 Ticino scans included. **robots.txt:** we found, audited and fixed our own robots.txt bug; every fetch path now fails closed (disclosure: `docs/incidents.md`).

5. Results without a model: schema verdict, census, repair

Table 1. The 17 fields. *Class:* role split = one label covers several roles, on OPD's labels or on the documents (P8); fragile = not under both P8 checks; untested = meaning not yet measured. *Source* of the split (committed results only): practice (parliaments differ; the standard encodes the role), OPD (likely a data or tagging error OPD can correct), both, or open (the labels cannot tell). *Where it lives:* parliaments where OPD metadata carries it · carries it for part of the affairs · only the PDF · not in OPD's data (`results/census/matrix.json`).

field	class	source	evidence	where it lives (meta · mixed · PDF · none)
answer date	role split	both	decided, dispatched or registered; in 7 of 54 bodies the submission date	metadata: 53 · 1 · 0 · 0
procedural history	role split	open	35 event labels, several date roles	metadata: 78 · 9 · 1 · 1
type	role split	both	42 of 1,038 raw labels map to two types; P8: the instrument	partly: 61 · 24 · 2 · 2
procedural state	role split	open	32 labels ambiguous	partly: 41 · 27 · 18 · 3
decision outcome	role split	practice	3 labels ambiguous	partly: 7 · 31 · 33 · 6
submitter	role split	practice; OPD in GR	P8: sole, joint or group author	partly: 23 · 61 · 5 · 0
co-signatories	role split	practice	P8: includes authors of record	PDF only: 0 · 2 · 72 · 12
language	role split	both	P8: tag names the other language; text, tag and affair differ	partly: 14 · 53 · 20 · 2
addressed body	split, fragile	practice	P8: the council (Bellinzona)	partly: 12 · 10 · 54 · 10
government recommendation	split, fragile	both	P8: by stage	partly: 3 · 8 · 56 · 10
submission date	body-specific in 2	OPD	P8: OPD's date printed in neither document (Vernier 5/5 affairs, Lausanne 3/5); a data-source split	metadata: 89 · 0 · 0 · 0

field	class	source	evidence	where it lives (meta · mixed · PDF · none)
answer-link status	untested	—	one v0.1 definition; meaning test next	metadata: 54 · 0 · 0 · 0
decision date	untested	—	one v0.1 definition; meaning test next	partly: 7 · 30 · 34 · 6
Q/A pairing	untested	—	one v0.1 definition; meaning test next	partly: 22 · 25 · 4 · 3
usable text	untested	—	one v0.1 definition; meaning test next	partly: 41 · 42 · 0 · 6
dispatch date	untested	—	one v0.1 definition; meaning test next	PDF only: 0 · 0 · 47 · 7
related affairs	untested	—	one v0.1 definition; meaning test next	PDF only: 0 · 0 · 76 · 13

Of the 10 role splits, 6 come from the role test (P8, registered) and 4 from OPD's label vocabularies (post hoc); by source of the split 4 parliamentary practice, 4 partly OPD's tagging (which OPD can correct), 2 open. 4 fields are in OPD's metadata almost everywhere, 13 partly or only in the PDF. Two further *concepts* are body-specific: oral answers in question time (23.6% of question-time answers) and councils that publish only meeting documents (16 bodies; v0.1 marks affairs rebuilt from minutes `virtual`, P4).

Cross-parliamentary question. Of the 91,268 interventions with a recorded government answer, 81.6% have answer text attached in OPD (the registered count); once the repairs apply, 85.6% with the repairs from what OPD already holds, 88.3% after the Bern relinks, 92.1% with the checked fetches (`results/report/r9_numbers.json`). *Answer times* (post hoc, 17 parliaments with at least 20 durations; figure in the README): on the same affairs OPD's answer dates and the documents' dates rank the parliaments almost identically (Spearman 0.997), so rankings do not hinge on the date's role; the role decides where OPD's answer event is the submission date: in Kriens the verified answers give 26.5 days on 68 affairs, OPD's date 0.

Where each schema field comes from, per parliament (41 of 89; all in `results/census/matrix.csv`)

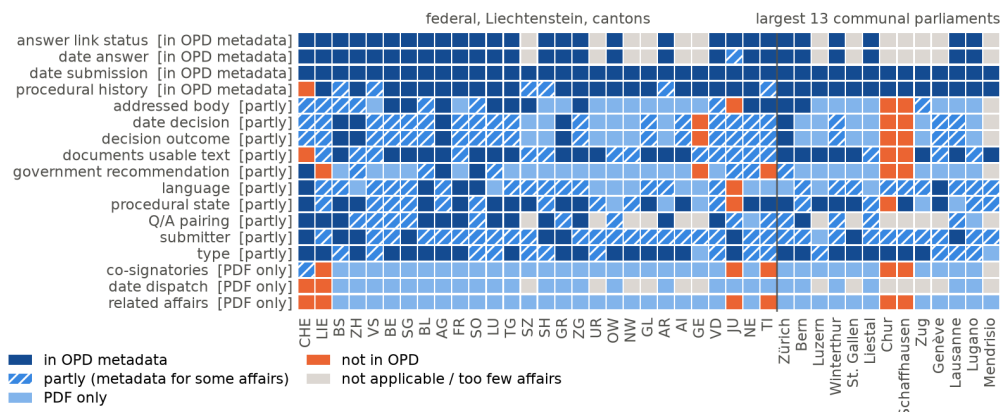


Figure 2. Where each field comes from, per parliament. Orange = not in OPD's data for that parliament (the parliament may still publish it).

Repair. Bern (a linking bug OPD confirmed; correspondence, 10 Oct 2026): of 3,532 answers the canton publishes OPD attaches 229 (6.5%), although 96.1% are in OPD and linkable without a model. 2,459 relink candidates = 2,421 that pass every check against OPD's own copy (affair number, heading, decision date = OPD's answer event [97.9, 98.9]) + 38 to review. Relinking changes each document's hash unless OPD keeps the old one; the plan carries both values. *Source adapters* (`docs/adapters.md`): readers for 50 parliaments (27 yield counted rows) verify 6,473 answers by affair number or title, the parliament's answer formula and a date check; 2,177 of them are already in OPD with text under a heading the census lexicon does not list: the repair is a category on the document (`tag_answer`). **Blind audit v3** (60 decoys, key hashed on GitHub first): on the counted rows the blind AI annotators judged 185/186 the executive's answer (Claude Opus) and Qwen-SEA-LION 186/186 (post hoc re-read over the strata that still count; as run on 200 rows: 197/200). Also for OPD: 9 harmonised vocabulary labels wrong as published; 7,876 of 133,215 answered interventions (5.9%, 73 parliaments) whose last OPD state still reads "submitted" although OPD holds the answer (`results/findings/stale_states.json`; the operator command flags it for review); and a Fedlex resolver for 29,946 federal placeholder documents.

6. Evaluation: ten pre-registrations

Design. Each test froze code, prompts and θ by sha256 before any label was committed (P1's hypothesis script followed its labels, D4; its rules came with its sample, D12). *Role split*: one label covers several roles; the field is shared once each value names its role. Unit (P1, P2, P6): one document's answer date, strict. Arms on identical inputs: **rules**; **pointer**; **cascade** (rules, then the pointer on the residue); **naive** free generation, plain and null-allowed; trivial baselines. **Gold, no human labelling by design**: two blind agent annotators (Claude Opus, Sonnet), a third breaks ties, quotes verified by code, raw files committed; all 96 annotator sessions (P1–P8, P10, the document-parts check, the federal cell) ended before GitHub received their labels

(results/eval/annotation_sessions.json). An isolation audit of their own transcripts lists every file the 96 sessions read and every command they ran: none touched a key, gold label, result, text outside its packets, another annotator's folder or the internet (0 violations; post hoc, results/eval/annotator_isolation.json). A manifest lists all 99 sessions (P11's 3 included) with start, end, model and the sha256 of the transcript and of each of the 1,393 label files (annotation_manifest.json); the transcripts are not committed because their packets contain document text.

Table 2. The ten tests, in registration order.

test	n	what it tests	registered result	note
P1	154 documents, parliaments with rules	answer date: rules vs Apertus	partial : H3 held; H1 null safety failed (7/90); H2 Branch B	rules alone 135/154, so no model by default
P2	62 documents, 15 parliaments without rules	rules transfer; pointer vs rules	3 of 4 held ; H3 null safety failed (4/33)	pointer 24/29 dates, rules 2/29
P3	29 whole affairs, 10 parliaments	7 fields: metadata, rules, Apertus	partial : no field unattended; Apertus recall held, precision below	feasibility study (D3-5)
P3b	36 fresh affairs, 9 parliaments	replication of P3	held : replicated (S1, S2)	S2 lenient; type 35/36; addressed body 2/33
P4	120 agenda items, 6 sittings	affairs rebuilt from minutes	partial : H2, H4 held; H1, H3 failed	recall 99/120, precision 99/99
P5	40 fresh affairs, 10 parliaments	Q/A pairs, people, recommendation	partial tier (H1); H2, H5 held; H4 as rules vs free generation; H3 no gain	<i>not yet</i> when parliaments are resampled
P6	68 fresh documents, 8 parliaments	answer-date reader v2 vs free prompting	no hypothesis held	free prompt 62/68, v2 59/68
P7	478 review rows, 190 blind packets	Apertus as verifier of the review queue	H1 holds (120/120); H4 supported; H3 not supported	214 of 917 rows promoted
P8	66 fresh affairs, 11 parliaments	meaning of 7 fields in the documents	none uniform : 6 role split, 1 body-specific	5 robust; meaning measured for 11 of 17
P10	34 fresh affairs, 14 parliaments without rules	author roles: Apertus reader vs rules, OPD's list	3 of 4 held : H1, H2, H3b; H3a not	pairs right 50/141, rules 17/141, OPD 37/141

Unattended (95% lower bound ≥ 0.80 , document-level intervals): the rules' answer date where rules exist (P1, helped by right nulls: dates found 47/64), the type on whole affairs (P3b, 35/36), and in P5 question detection, the reasoning span and co-signatory precision. **Cross-checks.** Qwen-SEA-LI0N-v4-32B-IT (AI Singapore) relabelled P1, P2, P6 and P7 blind: date κ 0.80, 0.87 and 0.86, P7's is-answer labels 185/190; no registered decision changes. Non-LLM checks: OPD's event matches 37 of 44 checkable P1 dates. Holm over all 20 registered confirmatory tests of P1–P7: the 9 that held survive (results/eval/posthoc_r8_multiplicity.json); whole parliaments are resampled as well.

P1 arm (154 documents)	exact [95% CI]	answered: found /64	wrong printed date	invented	answerless dated /90
rules only (default)	135/154 [81.5, 92]	47	4	0	2
pointer	130/154 [77.9, 89.3]	46	16	0	6
cascade	139/154 [84.6, 94]	56	11	0	7
naive, null-allowed	134/154 [80.8, 91.4]	47	11	0	3

P1: its rules entered with its sample (D12), so P2 is the test of transfer. H1 null safety failed (the cascade dates 7 of 90 answerless documents, 6 of them council reports; upper bound 15.2% vs 10%); H2 gave Branch B (cascade alone right on 5, rules alone on 1, $p = 0.22$), so Apertus left the unattended path for an opt-in review suggestion (D6); H3 held (AUROC 0.969).

P2 arm (62 documents)	exact [CI]	answered: found /29 [CI]	wrong date	invented	answerless dated /33
rules only	35/62 [44.1, 68.1]	2/29 [1.9, 22]	2	0	0
pointer	53/62 [74.7, 92.2]	24/29 [65.5, 92.4]	8	0	4
naive, null-allowed, re-parsed (D2-4)	54/62	24/29	7	1	3

P2: the rules collapse off their home layouts (47/64 $\rightarrow$ 2/29). H2 held: on all 62 documents the pointer beats the rules on 22 and loses on 4 ($p = 5.3 \times 10^{-4}$); resampling the 15 parliaments, the gain is [-0.02, 0.54], so it sits in a few parliaments. "Answerless" means no qualifying answer date: the pointer's 4 such dates fall on documents with an answer but no qualifying date; of the 25 with no answer at all it dates none (post hoc). By language (post hoc), the pointer finds answer dates in every language: DE 9/13, FR 10/11, IT 5/5.

P3/P3b/P4. Default (metadata + rules) on 29 affairs: type 27/29, answer date 22/27, 27/27 with Apertus suggestions. P3b replicated by its registered criteria; answer date 20/35 by default (C), 31/35 with Apertus suggestions. **P5.** Rules-only pairing: precision 0.83, recall 0.67, *partial* with document-level intervals, *not yet* when parliaments are resampled; recommendation 34/40 by rules; submitter rules 12/40, pointer 24/40, free generation 34/40.

P6 — reader v2 (an answer-present gate, then the guide's date rule). No registered hypothesis held: v2 vs free prompting 1:4 ($p = 0.38$); v2 dates 5/34 answerless documents (upper bound 30.1% vs 15%); v2 vs v1 5:2 ($p = 0.45$).

P7 — Apertus as a precision verifier of the review queue (eval/prereg8/, LOCK8 before the draw). Test frame: 478 eligible review rows outside every audit; 190 blind packets. H1 (primary): promotion precision 120/120, Wilson lower bound $96.9\% \geq 0.90$: **holds**. H4: dated precision 95.8% vs 73.2% for a trivial exact-date match, +22.7 points [10.2, 34.3], weighted, as registered (unweighted +7.5): **supported**. Base rate: 86.7% of eligible queue rows are answers, but only 40.9% are answers whose own date passes the rule, so the gain is the date check. The rule alone recovers more rows (369 vs 214 queue rows), at the lower dated precision above.

P8 — what the shared fields mean (eval/prereg9/, LOCK9 before the draw). Share in v0.1's meaning, pooled (parliament-cluster 95% interval): submitter 35/65 [33.3%, 75.0%], co-signatories 10/41 [6.8%, 46.3%], submission date 44/57 [57.1%, 91.8%], type 55/64 [76.2%, 95.3%], language 34/36 [83.3%, 100.0%], each verdict unchanged on agreed labels only and on Qwen's labels; addressed body (57/60) and recommendation (31/34) are fragile.

P10 — Apertus reads author roles where no rules exist (eval/prereg10/, LOCK10 before the draw and the labels; D1). 34 fresh affairs, 14 parliaments outside every rule set, 141 gold (name, role) pairs. Pairs right / printed: Apertus 50/96 (precision 0.52, recall 0.35, F1 0.42), rules 17/55 (0.31, 0.12, 0.17), OPD's contributor list as first author and co-signers 37/104 (0.36, 0.26, 0.30). Confirmed: H1, recall over the rules, +23.4 points [5.9, 49.0], Holm $p = 0.032$; H2, precision non-inferior (+0.21); H3b, over the same names with a trivial role. Not confirmed: H3a, over OPD's list (+0.12, Holm $p = 0.17$): where OPD lists the authors, its order is a strong baseline. F1 is 0.80 on P8's 6 development parliaments, where prompts and thresholds were chosen (in-sample), 0.42 on the 14 unseen ones; in Lugano (Italian, outside the family, the rules in-sample) 0.65, the rules 0/56 pairs.

Federal cell (exploratory, locked before draw and labels): 15 Federal Council answers print no answer date and no author; rules, pointer and converter date none (15/15 right nulls), the latest printed date dates an answerless document on 6/15.

P11 (exploratory, eval/p11/, LOCKP11 before draw and labels). *Italian*, 12 interventions from five parliaments: the pointer is exact on 12/12 [75.7, 100] (10 printed dates, 2 right nulls), the latest date on the answer document on 9/12, the rules on 8/12 (in-sample in three); author roles (70 gold pairs): the Apertus role reader F1 0.32, the converter's record 0.30, document order and OPD's list 0.22. *Outside interventions*, 15 government-business and initiative affairs: 7 of 17 fields carry over, 4 answer fields do not apply, 6 need new roles, chiefly the executive acting as proposer; the frozen converter returns a valid v0.2 record for 3 of 15 (registered arm), a writer from the same metadata and documents 15/15 (post hoc).

Threats to validity. The No-criterion (§1) was set after the results, and P8's frame codes a deviation that falls into a named category as a role split, so it leans toward "needs roles"; the answer to both is the role fit on gold labelled after v0.2 was frozen: every determinable value beyond authorship falls into a named role, 58/58 in 27 interventions from 6 parliaments, and 267/267 author roles; outside interventions 36/43 (exploratory, post hoc, oos_role_fit.json); the draft v0.2.1 adds the cited date role, which P8 already counts as shared (41/43, post hoc; v0.2 stays frozen). The same assistant family wrote the code and the gold labels; Qwen's blind relabelling changed no registered decision, and the isolation audit found no session that read a key, a result or another annotator's files. Three splits rest on one parliament each (§7).

Baselines proposed in review: docs/evaluation.md.

What the tests say. *Provenance by construction:* no arm that picks from printed dates returned a date printed nowhere (P1, P2, P6). **Null safety** is left to the reviewer: the registered null-safety tests on the 70B failed (P1-H1, P2-H3, P6-H2), which is why the default runs no model.

7. Next steps, reproducibility, declarations

Next steps. With OPD: its hash decision for the relinks and the 30-row spot-check kit; with the eCH group: QAItem and the five gaps. *Limits of the answer, each with its next test:* three splits rest on one parliament each (addressed body, recommendation, language): more parliaments per field. Italian rests on P11's exploratory cell (12 affairs; author-role F1 0.32 or lower for every arm): a registered Italian test. Outside interventions: v0.3 roles for the executive as proposer. 6 fields are untested: their meaning test. **Reproducibility.** make run (Docker, CPU) recomputes the census and the Bern join, re-derives the P1 rules arm from the source PDFs and P1's and P2's Apertus decisions from the model cache, replays P1–P8 and P10 (all locks verified), regenerates the summary-number files byte for byte and runs the claims check, which binds every number and scope word in this report to its claim and checks this PDF against its source. **Declarations.** Claude Code agents (Anthropic) wrote most code, documentation and the gold labels. Public AI (may route outside Switzerland) served evaluation only. Report CC-BY-4.0.