Inbrev

Sovereign inbox triage for Swiss medical practices

Niklas Wichter
Team Apertool · Hack Apertus 2026 · Track 2B — Own Project

- **Demo video (under 2 min):** `website/media/video.mp4` in the repository; public link in the submission form

Version of 16 October 2026. Every number is measured or cited; what is not measured yet is named as such. Data are synthetic; labels re-checked by one person (§4).

1. Summary

Swiss general practitioners spend more and more time on administration: 9.5 hours per week in 2025, up from 8.6 in 2020, and every additional administrative hour raises the odds of stress (Zeller et al., 2026). Physicians in private practice report 54 minutes per day for requests from authorities and insurers alone (Golder et al., 2025). Much of this arrives through the practice inbox (Arndt et al., 2017). Patient messages fall under professional secrecy, which binds physicians and their assistants alike (StGB, 1937), so sending them to a foreign cloud model is not an option.

Inbrev sorts the inbox of a Swiss practice locally with Apertus. It assigns each message to one of nine administrative routes, derives the administrative deadline and the person in charge, extracts patient and case fields, and backs every decision with a verbatim quote from the message. In its default mode it does not rate or rank clinical content: a message that may describe symptoms leaves automation and goes to a person, in the same queue as messages the model is unsure about, without any ranking by severity. Results (Apertus 8B; synthetic data, labels checked by one person on 60 messages):

- **Gold set, 180 messages, default configuration** (prompt p3 plus a second, narrow 8B check for current complaints, chosen on a dev split and then measured once here): all 16 messages with clinical content leave automation (keyword rules: 7), 170 of 180 routes are right, 140 of 180 messages run automatically and 40 go to a person. Before the second check: 13 of 16 and 151 automatic.
- **Clinical challenge set** (115 messages, 44 clinical, 39 medical traps, written after p3 was frozen; later part of the dev split for the second check): p3 alone excludes 41 of 44 with 8B, 44 of 44 with 70B, keyword rules 31 of 44.
- **Labels:** two open-weights models labelled all 367 test messages blind; they agree with our labels on clinical content (κ 0.97–0.98) and route (κ 0.93), and scoring against their labels does not change the ranking against the keyword rules. A person from the team re-labelled 60 blind: κ 0.98 route, 1.00 clinical (§4).
- **Other models:** Qwen3.5-9B and Ministral 8B in the identical pipeline route better than Apertus 8B (177 vs 170 of 180, wrong routes that run automatically 1 vs 8); §5 says why we still use Apertus. Clinical reports written by people are routed right (GraSCCo 40/40, PARHAF 37/38).
- **Offline and cost:** with local Q4 weights Inbrev runs on a Docker network without internet access (logged Docker run on a laptop CPU before the symptom pass: 192 s per message; natively on its integrated GPU 62.7 s, 71.8 s with the symptom pass); a mini PC costs about CHF 12 per 1,000 messages at 60 a day.

2. Architecture

Inbrev is a fixed pipeline, not an agent. Small local models are reliable on narrowly cut tasks and weak on complex ones (Xu et al., 2026), so the model does only what rules cannot: read free text in four languages and point to the passage that supports a decision.

1. **Intake.** E-mail (`.eml` or IMAP, read-only, including PDF attachments), HIN mail through an IMAP bridge (untested), fax pages as PDF or image (Tesseract OCR for German, French and Italian), web forms.
2. **Apertus with an enforced JSON schema** (Geng et al., 2023; Willard & Louf, 2023). The schema orders each quote before its decision (route quote → route, safety quote → flag, deadline quote → date). Three samples are drawn in one request.
3. **Deterministic checks per sample.** Every quote must occur in the message: verbatim, after case and whitespace folding, or at least 90 % similar to one span; the evidence shown to a person is always the original text, never the model's copy. Language models often cite incompletely (Gao et al., 2023), and Apertus showed low faithfulness in production use (Kruckenberg, 2026), so a value without a verifiable quote is dropped. Dates must match their quote; AHV numbers must occur in the text; fields that do not belong to the chosen route are removed.
4. **Rules instead of model judgement where possible.** The deadline category is computed in code from the quoted date and the time of receipt; without a quoted date or time there is no deadline, except that flagged messages and received reports are due today. The assignee follows from the route. A safety flag always means "doctor, today".
5. **Vote and abstention.** Route and fields by majority over the samples; the safety flag is raised if any sample or a multilingual lexical net fires. Agreement across samples is used as the uncertainty signal instead of a verbalised confidence, which is overconfident (Wang et al., 2023; Xiong et al., 2024). A second, short 8B call (default since 13 October) asks only whether someone describes a current complaint and must quote it; one verified yes out of three samples also raises the flag (not for reports, which go to the doctor anyway); if this call fails or returns no answer, a person decides. If the samples disagree, the model marks the message as ambiguous,

Phase	Dependencies
Build time	Docker images (Python, llama.cpp), Python packages, Tesseract language data, Apertus weights from Hugging Face (checksum)
Runtime	none; the model is reached over an internal Docker network without internet access. Optional: IMAP mail pick-up

or no route evidence survives, Inbrev abstains: a person decides (Wen et al., 2025).

6. **Output.** A sorted inbox (web UI: "a person decides" for unsure messages and messages with clinical content, by time of receipt and without a ranking among them; then today, this week, no deadline), a JSON result per message following `data/schema/triage_result.schema.json`, an API under `/v1/`, and an audit log that stores a hash and the decision, not the text. Reply drafts are fixed templates for administrative routes only; no model-written text reaches patients.

Target architecture (mandatory)
Inbrev is built for **a) on-premise** and **b) air-gapped** operation; **c) a sovereign Swiss cloud** is an option for single-doctor practices.

Health data are sensitive personal data under the Swiss Data Protection Act, which also requires data protection by design and allows processing by third parties only where no duty of secrecy forbids it (DSG, 2020). A system whose data never leave the practice avoids the question. Open weights that run on the practice's own hardware also keep the model auditable (Project Apertus et al., 2025; Riedemann et al., 2024).

Offline run with local weights. `make offline-eval` (`scripts/run_offline_eval.sh`) starts llama.cpp with a community Q4_K_M GGUF of the Apertus 8B text weights (Apache-2.0) and Inbrev on a Docker network with `internal: true`; only an optional nginx gateway for the web UI also sits on the host network. The log records the SHA-256 of the loaded weights (equal to the Hugging Face LFS hash, revision `248ec68`), `docker inspect` of both containers, an egress test from the Inbrev container (DNS, TCP to 1.1.1.1:443 and HTTPS all fail) and a control on the default network (all succeed), memory samples and the evaluation. On a 28-W laptop CPU (Ryzen 7 PRO 8840U, Docker Desktop) the model needs 6.2 GiB RAM and Inbrev 167 MiB; results are in §5. The run is in `evidence/offline/`.

3. Use of Apertus

- **Model:** swiss-ai/Apertus-v1.5-8B for inference; Apertus-v1.5-70B as a comparison, recommended where the hardware allows it (§5) (Project Apertus et al., 2025)
- **How it is used:** inference (classification and extraction with an enforced JSON schema, three samples per message; one on a CPU)
- **Where it runs:** development and evaluation on the Hack Apertus endpoint (OpenAI-compatible, vLLM behind a LiteLLM proxy); target deployment on local weights (llama.cpp, GGUF Q4) inside the practice, measured offline on a CPU (§2, §5)

Settings: temperature 0.7 for the three samples, at most 900 output tokens, thinking off because it lowers instruction following (Wiegand, 2026); we did not measure thinking mode ourselves. The system prompt (p2; p3 since 4 October, see section 5) lists the routes with short multilingual cues, the rules for safety, deadlines and quotes, and an output format with placeholders only.

4. Data

All messages, people, practices and identifiers are fictional; AHV numbers have a valid format and an invalid check digit. No real patient data is used. Licence: CDLA-Permissive-2.0.

- **Seed:** 12 messages (German, French, Italian; e-mail, fax, web form) with gold labels, in the repository. Claude (`claude-opus-5-5`) wrote them one by one in an agent session, not a person; they served as the development set for the prompt and the checks.
- **Synthetic part (v0):** 180 messages. Labels are fixed by a seeded grid *before* any text exists; two generator families write the text from one documented prompt, 90 messages each (Claude and GPT-5.5). Apertus neither writes nor labels data (Panickssery et al., 2024). German 99, French 45, Italian 27, English 9; 16 messages with clinical content (7 explicit, 9 in passing), 14 medical traps, 9 ambiguous messages. A validator checks schema, verbatim quotes, dates, invalid AHV check digits, fictitious domains and near-duplicates.
- **Swiss practice set:** 30 messages (`data/ood/`), text and labels by an Antigravity agent session (Gemini, `gemini-3.8-flash`) on 3 October; real organisation names and domains were later replaced by fictitious ones (`scripts/annotate_swiss_set.py`); each row records its edits in `provenance`; the generation prompt was not recorded.
- **Open-generator set:** 42 messages from a fresh seeded grid (seed 2027, 8 with clinical content, 7 traps), text by an open-weights model, NVIDIA Nemotron 3 Ultra (NVIDIA, 2026), through OpenRouter with the same prompt minus all example values; 3 of 45 rows failed validation and were dropped. Nobody tuned prompt, checks or lexicon on it. It tests whether the results depend on the closed generators.
- **GraSCCo report set:** 40 of the 63 synthetic German clinical letters of GraSCCo v1.1 (CC BY 4.0 on Zenodo, the authors also state CC0) (Modersohn et al., 2022), verbatim (5 shortened at a paragraph end) behind a fixed one-line cover note, as realistic long `report_received` messages (`data/grascco/`). Labels follow one rule: route `report_received`, deadline today, no clinical flag. The letters are German and Austrian, not Swiss.
- **Clinical-exclusion challenge set:** 115 messages from a new seeded grid (seed 5151), 44 with clinical content (17 explicit, 27 in passing), 39 medical traps (negations, past events, chronic conditions), more typos and dialect; text by Nemotron 3 Ultra (open weights, free tier), written on 13 October after prompt, checks and lexical net were frozen (`evidence/challenge/`). It was unseen for p3, but

it is part of the dev split on which the second check was chosen, so it is not a hold-out for that check.
- **PARHAF report set:** 38 French clinical reports from PARHAF (fictitious cases written by resident physicians; CC BY 4.0 / Etalab 2.0), verbatim behind a fixed cover note, same rule as GraSCCo (`data/parhaf/`).
- **External datasets** (Hugging Face) served only local stress tests; they are not redistributed and give no reported number (`data/README.md`).
- **Label check by two models and one person:** two open-weights models, Nemotron 3 Super and gpt-oss-120b, labelled all 367 test messages blind from the written rules, seeing only the message (`scripts/annotate_open_models.py`, USD 0.06). Their agreement shows that the rules can be applied consistently, not that labels are right. One person from the team (not a clinician) re-labelled 60 gold messages blind: Cohen's κ (Cohen, 1960) 0.98 [0.94, 1.00] route, 1.00 deadline, 1.00 clinical. Of 36 disputed messages, our label was right in 27, the model in 9; no label was changed (`evidence/review/`).

Closed models (Claude, GPT-5.5 via Codex, Gemini via Antigravity) wrote or edited test data and code during development only, and reviewed drafts. Open-weights models have three supporting roles: Nemotron 3 Ultra wrote the Nemotron, dev and challenge sets; Nemotron 3 Super and gpt-oss-120b are the second annotators; Qwen3.5-9B and Ministral 8B run once in the Inbrev pipeline as a comparison (Llama 3.1 8B was tried, but the provider rate-limited it). At runtime Inbrev calls only Apertus.

Synthetic data can miss the distribution of real mail (Long et al., 2024); only the GraSCCo and PARHAF reports were written by people, and they cover a single route.

5. Evaluation

The main numbers are the clinical exclusion recall (share of messages with clinical content that the safety net takes out of automation; "safety" in the tables), the recall of the deadline "today" and the share of messages that run without a person (coverage). Clinical messages are rare, so recall and precision are reported instead of accuracy (Saito & Rehmsmeier, 2015). Further metrics: route macro-F1, deadline accuracy, field F1, evidence faithfulness (Rashkin et al., 2023), tokens, latency and cost.

Results on the synthetic set v0 (n = 180; 16 clinical cases, 50 "today" deadlines). TF-IDF uses five-fold cross-validation. Inbrev rows run the full pipeline (schema, three samples at T = 0.7, checks, vote) on the Hack Apertus endpoint, p2 on 3 October, p3 on 4 October, p3 with the second check (default) on 13 October. Coverage counts messages that are neither abstained nor flagged, at most 164 of 180. "One sample, no checks" scores the first stored 8B sample as written, without checks, net, vote or abstention (`scripts/ablation_first_sample.py`).

With 95 % Wilson intervals (p2): clinical exclusion 13/16 [0.57, 0.93] for 8B and 15/16 [0.72, 0.99] for 70B (McNemar p = 0.50), against 9/16 [0.33, 0.77] for a single 8B answer without checks and 7/16 [0.23, 0.67] for the keyword rules. Three samples add two cases (11/16 without the lexical net), the net two more; it is tuned for recall, and about half of the flags on this set are false alarms from the net (15 of 28 for 8B), which costs a doctor a glance, not a missed case. Recall differs by generator (8B: 7/7 for texts by GPT-5.5, 6/9 by Claude), so the set partly measures its authors; the open-generator and challenge sets are the control. Field F1 is 0.73 for 8B and 0.83 for 70B; mean latency per message 3.0 s and 10.7 s on the shared endpoint.

The baselines already route 79–83 % of the synthetic messages correctly, so the route alone does not justify a language model on this set. They fail where it matters: they find at most 44 % of the messages with clinical content.

Where Apertus beats the rules. Paired on the 180 messages, 8B is right where the rules are wrong 27 times and the reverse 11 times with p2 (McNemar p = 0.014), 32 and 5 times with p3 ($p = 7{\cdot}10^{-6}$): it wins on insurer mail, marketing and symptoms mentioned in passing. A rules-first hybrid (rules decide when exactly one route family matches and the net is silent) needs the model for 94 of 180 messages but drops to clinical exclusion 10/16, because 5 of 16 clinical messages look like plain administration to the rules; on the challenge set it drops from 41 to 36 of 44. We therefore keep Apertus on every message (`scripts/compare_hybrid.py`).

Challenge set. The 44 clinical messages narrow the interval of the main metric: 8B 41/44, against 31/44 for the keyword rules (10 vs 0 discordant, exact McNemar p = 0.002), 70B 44/44 (8B vs 70B p = 0.25). Of 71 non-clinical messages, including 39 traps, 8B and 70B each flag 4 (precision 0.91 and 0.92). Route: 8B wins 20 vs 4 against the rules (p = 0.0015). A TF-IDF classifier cross-validated on the set itself reaches 39/44 (an in-domain reference, not deployable). **Reports.** On GraSCCo and PARHAF the false flags come almost only from the lexical net (6 of 7; 16 of 17, mostly "urgence" in emergency department reports); Apertus itself flags 1 of 38 French reports. Each costs a doctor's glance at a report that goes to the doctor anyway; we did not retune the net on these test sets. **Fax.** 60 gold messages rendered as degraded fax pages and read by Inbrev's own Tesseract

Setup	Route acc.	Route F1	Safety rec.	Safety prec.	Today rec.	Deadl. acc.	Coverage
Majority class	0.14	0.03	0.00	–	0.00	0.37	1.00
Keyword rules + lexical net	0.79	0.76	0.44	0.32	0.76	0.68	0.88
TF-IDF + logistic regression	0.83	0.82	0.13	1.00	0.66	0.68	0.99
8B, one sample, no checks	0.87	0.84	0.56	1.00	0.70	0.67	0.95
Inbrev 8B, p2 (schema, checks, vote)	0.88	0.85	0.81	0.46	0.86	0.82	0.81
Inbrev 70B, p2	0.91	0.89	0.94	0.50	0.94	0.85	0.79
Inbrev 8B, p3	0.94	0.94	0.81	0.46	0.88	0.84	0.84
Inbrev 8B, p3 + second check (default)	0.94	0.94	1.00	0.41	0.94	0.81	0.78
Inbrev 70B, p3	0.96	0.96	0.94	0.50	0.94	0.87	0.82

Further sets (Inbrev 8B, p3 unless noted)	n	Route acc.	Clinical excl.	Keyword rules route / clinical
Challenge set (8B, written after p3)	115	103/115 [0.83, 0.94]	**41/44 [0.82, 0.98]**	87/115 / 31/44
Challenge set, 70B	115	106/115 [0.86, 0.96]	**44/44 [0.92, 1.00]**	–
Open generator (Nemotron)	42	39/42 [0.81, 0.98]	8/8 [0.68, 1.00]	32/42 / 6/8
GraSCCo letters (German, p2)	40	40/40 [0.91, 1.00]	– (7 false flags)	40/40 / –
PARHAF reports (French)	38	37/38 [0.87, 1.00]	– (17 false flags)	3/38 / –
Rendered fax pages via OCR (gold subset)	60	50/60 [0.72, 0.91]	12/16 flagged, 13/16 out of automation	–

intake (character error rate 0.27, upper bound): the route falls from 56 to 50 of 60, the 3 unreadable pages abstain, and 13 of 16 clinical messages still leave automation, as on clean text (`evidence/fax/`).

Label source. Two open-weights models labelled all 367 messages of the four generated sets blind (§4). Cohen's κ with our grid labels is 0.93 for the route and 0.97–0.98 for clinical content, 0.60–0.90 for the deadline and 0.40–0.42 for "ambiguous". Scored against either model or the majority of three, Inbrev 8B routes 167–169 of 180 gold messages (170 against the grid) and keeps 13/16; on every set the ranking against the keyword rules is unchanged (`evidence/agreement/`).

Other open models, and why Apertus anyway. In the identical pipeline (prompt p3, checks, three samples) on the gold set, Qwen3.5-9B and Ministral 8B both route 177/180 and exclude 16/16 clinical messages, against 170/180 and 13/16 for Apertus 8B (route p = 0.04, clinical p = 0.25) and 172/180 and 15/16 for Apertus 70B (`evidence/peers/`), although the prompt was developed with Apertus. With prompt p3 alone, errors that pass automation unnoticed (wrong route, or clinical content without a flag) are 9 + 3 for Apertus 8B, 8 + 1 for 70B and 1 + 0 for both comparison models. We then tried to close this gap with Apertus only, choosing on a dev split (175 messages, rule fixed before the hold-out run, `evidence/gap/`): a safety-first schema order (worse), five samples (no effect) and few-shot examples (partial) were dropped; the symptom pass above reached 0 unnoticed clinical messages on dev. Measured once on the gold set, Apertus 8B with the symptom pass lets through 8 + 0 at 40 messages for a person (Qwen 1 + 0 at 38); adding a gate that sends a message to a person when keyword rules choose another route gives 4 + 0 at 69. The clinical part of the gap is closed, the route part only narrowed, and only by more human work. **Deviation from the fixed rule:** the rule, written before the hold-out run, picked the second check plus the rules gate; after seeing the hold-out we ship the second check alone as default and the gate as an option (`INBREV_RULES_GATE=1`), because the gate sends 69 instead of 40 of 180 messages to a person. A second round on a fresh hold-out (146 messages written by Gemma 4 31B, frozen with hashes before any run, plan fixed on the earlier sets; `evidence/route/`) targeted the routes: sharper route descriptions and few-shot examples alone change nothing (5 wrong routes run automatically, 39–42 to a person); with them (`INBREV_PROMPT=p4`, `INBREV_PROMPT_VARIANT=fewshot4`), asking 70B when the keyword rules disagree lowers this to 2 at 43 (3 at 45 with the default prompt), available as an option for practices with 70B access (`INBREV_SECOND_OPINION_MODEL`, off by default). Qwen3.5-9B: 0 at 24. Local execution, the quote checks and the unranked safety net work with any open model and are no argument for Apertus. What is specific to Apertus: it was pretrained only on openly available data with robots.txt opt-outs applied retroactively and personal data filtered, all training artefacts are released (Project Apertus et al., 2025), and the developers publish a dataset catalogue, an EU training-content summary and an EU Code of Practice statement (github.com/swiss-ai/apertus-data-transparency), the kind of documentation the EU AI Act asks of general-purpose model providers (EU AI Act, 2024); we did not check equivalent documents for the comparison models. A Swiss practice that must justify under the DSG and professional secrecy which model reads patient mail (DSG, 2020) can point to this provenance. That is a trade-off, not a free choice: the measured price is 7 more wrong routes in 180 that reach automation, or more human work to catch them. Inbrev talks to any OpenAI-compatible endpoint, so a practice can choose otherwise; with Apertus we recommend 70B where the hardware allows it.

Prompt p3 for "other". p3 only sharpens route boundaries ("other": copy of records, referral elsewhere, symptoms only). It was developed on a separate dev set (60 Nemotron messages; "other" p2 0/20, p3 11/20), frozen after one dev run, then run once on every test set: gold route 158 ⟶ 170/180 (McNemar p = 0.002), no clinical case lost (`evidence/prompt-p3/`).

One sample is nearly as good as three. Replaying the stored 8B run (p2) with only the first sample gives route 157/180 and 12/16 against 159/180 and 13/16 in the stored run of 3 October (`scripts/replay_k_samples.py`; the shipped p2 rerun `gold-p2-8b.jsonl` gives 158/180, hence the 158 baseline for p3 above, see `evidence/prompt-p3/noise-p2-rerun.md`), so local deployments use one sample.

Offline on a CPU. On 30 of the 180 messages (all 16 clinical plus 14), the offline Docker stack with Q4 weights and one sample (p2) routes 24/30 and excludes 11/16; the endpoint (BF16, one sample) 23/30 and 12/16; Q4 and BF16 agree on the route for 28 and on the flag for 29 of 30. Wall time was 192 s per message on a laptop CPU. The same build and weights natively on the laptop's integrated GPU (Vulkan, `scripts/run_native_llamacpp.ps1`) need 62.7 s per message (p95 85 s); on the challenge set this configuration (Q4, one sample, without the symptom pass) excludes

38 of 44 clinical messages against 40/44 for one BF16 sample (`evidence/native/`). These native runs measure speed and quality; the air-gap proof remains the Docker run. With the symptom pass (default) the same native setup needs 71.8 s per message and excludes 16/16 on these 30 messages (route 25/30).

Cost. From measured time and cited prices (excl. VAT; `scripts/cost_model.py`): a mini PC with the same CPU draws 38–39 W under load (Kennedy, 2024), about 2.1 kWh or CHF 0.52 per 1,000 messages at 192 s each (ElCom, 2026). With the box (about CHF 700, three years) and idle power, a practice with 60 messages a day pays about CHF 12 per 1,000 messages, roughly CHF 22 a month; at 71.8 s per message (iGPU, default), 60 messages a day are about 72 minutes of machine time, so sorting can run as mail arrives. A Swiss token API at the listed Apertus 70B price costs CHF 3.13 per 1,000 messages in the default configuration (tokens measured with the symptom pass, `evidence/value/cost-default-config.md`). **Value.** Staff time saved is not measured. `scripts/value_model.py` multiplies measured shares of the default configuration (78 % automatic, 4.4 % automatic but wrongly routed, 12.8 % false clinical flags) with assumed volumes and handling times (`evidence/value/`). At 60 messages a day, 60 s by hand and 15 s to check a suggestion, an MPA saves about 32 minutes a day, net CHF 349 a month after doctor time for false flags and running cost (all assumed); the cautious case loses CHF 242 a month; break-even is 33 s of manual handling per message. Practice interviews with a timing trial will replace the assumptions.

Not measured yet: latency on an NVIDIA GPU (`scripts/run_native_llamacpp.ps1 -Backend cuda` is prepared) and the time to sort messages by hand versus with Inbrev.

6. Limitations

- The synthetic set partly measures its generators; the strong surface baselines suggest it is easier to separate than real mail. Messages written by people (practice staff or patients) would be the stronger test; none are in the evaluation yet (Long et al., 2024; Panickssery et al., 2024).
- Labels were re-checked only by one person from the team (not a clinician), on a sample of 60 (§4). The second labels come from two open-weights models (§4, §5), which measure whether the written rules are applied consistently, not clinical correctness; one is from the same family as the generator of two sets.
- **Regulatory boundary.** Software that automatically reprioritises by clinical criticality based on patient data qualifies as medical device software (MDCG, 2025); in Switzerland classification follows Annex VIII of the EU MDR (EU MDR, 2017; MepV, 2020). Inbrev therefore never orders by severity in its default mode: the safety flag only takes a message out of automation and puts it in the same "a person decides" queue as unsure messages, ordered by time of receipt. Intended purpose: sorting administrative mail; messages with clinical content are not processed automatically. The final classification depends on the written intended purpose; this is not legal advice.
- **Optional research mode.** `--mode clinical-triage` (off by default) adds a keyword-based urgency hint to flagged messages: no intended medical purpose, not validated, not certified; clinical use would need an MDR/MDSW conformity assessment. All numbers use the default mode.
- No measurement on real practice data; the GraSCCo letters are German and Austrian, the Nemotron set reuses our grid and prompt.

7. Reproducibility

`make run` starts the web app, `make run-local` the offline stack, `make offline-eval` reruns the offline pipeline (command for the logged run: `evidence/offline/`), `make test` runs the 35 unit tests with `--network none`. Scripts in `scripts/` recompute agreement, value, rules comparison, ablation and cost from stored runs; `python -m inbrev triage|replay|eval` reproduces the numbers from the stored runs, with 95 % Wilson intervals. Grids and cross-validation are seeded. Model runs use `apertus-v1.5-8b` and `apertus-v1.5-70b` as named by the Hack Apertus endpoint (3, 4 and 13 October), which does not expose the Hugging Face revision. The generated test texts are in `data/synthetic/` (synthetic, labels unreviewed); predictions, grids and annotations are in `evidence/`.

8. Next steps

- Five to ten interviews with practice assistants and practice owners: real volume, category mix, share of fax; a pilot in a group practice.
- HIN connection (the IMAP import is tested against a local test server only) and practice software; a data protection impact assessment template (DSG, 2020).

License

Creative Commons Attribution 4.0 (CC-BY-4.0). All Hack-Apertus projects are open-sourced. Code: Apache-2.0. Data (`data/`, including the generated test sets): CDLA-Permissive-2.0, except the GraSCCo and PARHAF letters (CC BY 4.0). External datasets used only for local checks are not redistributed and keep their own licences.

References

Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W.-J., Sinsky, C. A., & Gilchrist, V. J. (2017). Tethered to the EHR: Primary Care Physician Workload Assessment Using EHR Event Log Data and Time-Motion Observations. *The Annals of Family Medicine*, *15*(5), 419–426. https://doi.org/10.1370/afm.2121

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

DSG. (2020,). *Bundesgesetz vom 25. September 2020 über den Datenschutz (Datenschutzgesetz, DSG), SR 235.1*. https://www.fedlex.admin.ch/eli/cc/2022/491/de

ElCom. (2026,). *Strompreise 2026, Medianpreis Kategorie C2*. https://www.strompreis.elcom.admin.ch/map?period=2026&category=C2&priceComponent=total&product=standard

EU AI Act. (2024,). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying*

down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj

EU MDR. (2017,). *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices*. https://eur-lex.europa.eu/eli/reg/2017/745/oj

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 6465–6488. https://doi.org/10.18653/v1/2023.emnlp-main.398

Geng, S., Josifoski, M., Peyrard, M., & West, R. (2023). Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10932–10952. https://doi.org/10.18653/v1/2023.emnlp-main.674

Golder, L., Jans, C., & Schäfer, S. (2025,). *Befragung zum ärztlichen Arbeitsumfeld im Auftrag der FMH 2025*. https://cockpit.gfsbern.ch/de/cockpit/befragung-zum-aerztlichen-arbeitsumfeld-im-auftrag-der-fmh-2025

Kennedy, P. (2024,). *ASRock Industrial 4x4 BOX-8840U Review*. https://www.servethehome.com/asrock-industrial-4x4-box-8840u-review-low-power-amd-ryzen-mini-pc/3/

Kruckenberg, J. (2026, July 9). *Apertus after 8 months – what we learned and look forward to using Switzerland's AI Model*. https://www.liip.ch/en/blog/apertus-after-8-months-what-we-learned-and-look-forward-using-switzerlands-ai-model

Long, L., Wang, R., Xiao, R., Zhao, J., Ding, X., Chen, G., & Wang, H. (2024). On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. *Findings of the Association for Computational Linguistics ACL 2024*, 11065–11082. https://doi.org/10.18653/v1/2024.findings-acl.658

MDCG. (2025, June). *MDCG 2019-11 Rev.1: Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745 – MDR and Regulation (EU) 2017/746 – IVDR*. https://health.ec.europa.eu/document/download/b45335c5-1679-4c71-a91c-fc7a4d37f12b_en?filename=mdcg_2019_11_en.pdf

MepV. (2020,). *Medizinprodukteverordnung vom 1. Juli 2020 (MepV), SR 812.213*. https://www.fedlex.admin.ch/eli/cc/2020/552/de

Modersohn, L., Schulz, S., Lohr, C., & Hahn, U. (2022). GRASCCO – The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. *Studies in Health Technology and Informatics*, *296*, 66–72. https://doi.org/10.3233/SHTI220805

NVIDIA. (2026,). *NVIDIA Nemotron 3 Ultra 550B-A55B*. https://huggingface.co/nvidia/NVIDIA-Nemotron-3-Ultra-550B-A55B-BF16

Panickssery, A., Bowman, S., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. *Advances in Neural Information Processing Systems 37*, 68772–68802. https://doi.org/10.52202/079017-2197

Project Apertus, Hernández-Cano, A., Hägele, A., & others. (2025,). *Apertus: Democratizing Open and Compliant LLMs for Global Language Environments*. https://doi.org/10.48550/arXiv.2509.14233

Rashkin, H., Nikolaev, V., Lamm, M., Aroyo, L., Collins, M., Das, D., Petrov, S., Tomar, G. S., Turc, I., & Reitter, D. (2023). Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, *49*(4), 777–840. https://doi.org/10.1162/coli_a_00486

Riedemann, L., Labonne, M., & Gilbert, S. (2024). The path forward for large language models in medicine is open. *Npj Digital Medicine*, *7*(1), 339. https://doi.org/10.1038/s41746-024-01344-w

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

StGB. (1937,). *Schweizerisches Strafgesetzbuch vom 21. Dezember 1937, SR 311.0, Art. 321: Verletzung des Berufsgeheimnisses*. https://www.fedlex.admin.ch/eli/cc/54/757_781_799/de#art_321

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (2023,). Self-Consistency Improves Chain of Thought Reasoning in Language Models. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2203.11171

Wen, B., Yao, J., Feng, S., Xu, C., Tsvetkov, Y., Howe, B., & Wang, L. L. (2025). Know Your Limits: A Survey of Abstention in Large Language Models. *Transactions of the Association for Computational Linguistics*, *13*, 529–556. https://doi.org/10.1162/tacl_a_00754

Wiegand, G.-H. (2026, July 29). *LLM Benchmark Evaluation – Apertus 1.5-8B*. https://blog.nlp-lab.ai/2026/07/29/Apertus15Bench.html

Willard, B. T., & Louf, R. (2023,). *Efficient Guided Generation for Large Language Models*. https://doi.org/10.48550/arXiv.2307.09702

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024,). Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2306.13063

Xu, H. A., Pythoud, R., Thorball, C. W., Carra, G., Kulynych, B., Despraz, J., Galland-Decker, C., Maslias, E., Baudson, E., Brahier, T., Kraege, V., Sousa Teixeira, A. C. de, Fidalgo, C., Berthaudin, F., Madina, A., Zoergiebel, S., Bastardot, F., Stravodimou, A., Méan, M., … Raisaro, J. L. (2026). "Small" Large Language Models in the Hospital: Evaluation Study on Real-World Data in a Resource-Constrained Setting. *Journal of Medical Internet Research*, *28*, e86453. https://doi.org/10.2196/86453

Zeller, A., Fischer, R., & Hernandez, L. D. (2026). Between workload, job satisfaction and stress in Swiss primary care – results from the 5th Workforce Study 2025. *Praxis*, *115*(3), 67–72. https://doi.org/10.23785/PRAXIS.2026.03.002