Apertool

HACK APERTUS 2026 · TEAM APERTOOL

Can you trust Apertus with Swiss everyday life?



Our answer: yes, with the page in front of it. Not from memory.

Four tools on Apertus, the open Swiss model, each with its evidence.

Niklas Wichter · solo builder · **apertool.ch**

Apertool is an independent Hack Apertus team, not part of Apertus or its developers (EPFL, ETH Zurich, CSCS).

THE IDEA

One handle. Four blades.

Apertool is our name for the four tools. Apertus is the model they run on.



CheeseDrill Track 1A
Corkscrew · **tests** the model

Inkraft Track 1B
Saw · **measures** it

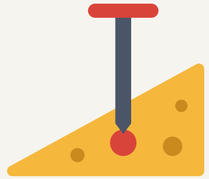
GeaNaVid Track 2A
Loupe · **checks** a claim

Inbrev Track 2B
Knife · **opens** the mail

One tool tests the model, one measures it, two put it to work.

All four on Apertus v1.5, one in each Hack Apertus challenge.

BLADE 1 · CORKSCREW · TRACK 1A RED TEAMING



CheeseDrill

We drill where Apertus made promises.

Project page opens on 1 December 2026



PROBLEM

The Swiss AI Charter sets out how Apertus should behave. We test where those promises give way on everyday Swiss questions.

BUILT

- Paired probes in German, French and Italian: each has a twin that changes one thing.
- One command replays every recorded run, no API key.

1 Dec 2026

results go public

Results are reported to the Apertus team first. Until then, CheeseDrill shows how it tests, not what it found.

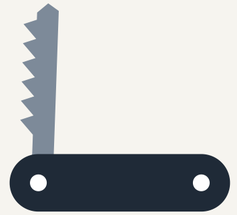
Track rule: results stay private until 1 December 2026. Technical docs in the README.

BLADE 2 · SAW · TRACK 1B SWISS VOICES

Inkraft

The law is the answer key.

inkraft.apertool.ch



PROBLEM

Social insurance law changes on fixed dates. A model can't tell which version it learned.

Minimum full AHV pension, per month

~~CHF 1,225~~ → **CHF 1,260**

until Dec 2024 — since Jan 2025 · Art. 34 para. 5 AHVG

BUILT

- A benchmark in German, French and Italian. Every gold answer rests on a quote from fedlex.
- It tells correct, stale and invented answers apart.

7 %

right on changed rules

Apertus 70B, 157 rows with an amended rule. **It gives the old law in 30 %.**

Majority of three open judges (κ 0.95). A human graded a blind 22 % sample: κ 0.91.

Inkraft v1.0.0: 326 questions, 948 rows in DE, FR and IT, drafted by LLMs, quoted from fedlex.

BLADE 3 · LOUPE · TRACK 2A CHALLENGE BY OST

GeaNaVid

Gea · Na · Vid: Romansh for yes · no · blank.

geanavid.apertool.ch · live demo: geanavid.apertool.ch/demo

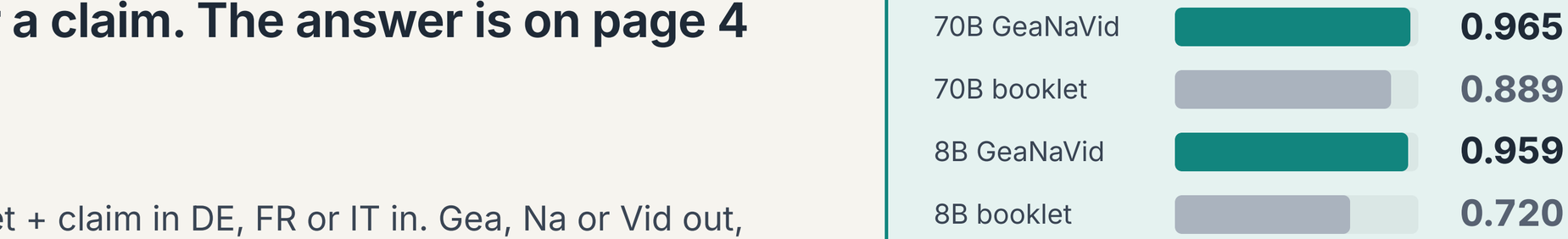
PROBLEM

Voters hear a claim. The answer is on page 4 of 40.

BUILT

- Voting booklet + claim in DE, FR or IT in. Gea, Na or Vid out, with the booklet sentence behind it.
- It reads the five best passages, not the whole booklet. Never voting advice.

Macro-F1, OST test split



20,242 tokens per claim for the whole booklet. **2,159** with GeaNaVid (8B), **1,461** (70B).

OST v1.1, 493 claims. Unseen booklets: 0.874 (8B), 0.932 (70B). The submitted default is 70B.

BLADE 4 · KNIFE · TRACK 2B OWN PROJECT

Inbrev

Your practice inbox, sorted where it belongs: in the practice, or in the Swiss cloud.

inbrev.apertool.ch



PROBLEM

Before the first patient, there is the inbox. Swiss GPs spend 9.5 h a week on admin, all under professional secrecy.

BUILT

- Sorts every message by handling path and deadline. Every field quotes its source.
- Anything that may need a clinician goes to a person, unranked.

16/16

clinical cases to a person

Keyword rules catch 7 of 16.

MEASURED 180 synthetic messages, labels from two open models, 60 checked by one person (κ 0.98 route). Route macro-F1: 0.94 (8B), 0.96 (70B), rules 0.76.

8B with second check · Zeller et al., Praxis 2026 (n = 1,776) · Art. 321 Swiss Criminal Code.

SO, CAN YOU TRUST IT?

Yes, with the page in front of it. Not from memory.

Inkraft · 70B

7 % → 99 %

right on the 157 amended rules: from memory, then with the law text in the prompt

GeaNaVid · 8B

0.72 → 0.96

macro-F1: the whole booklet, then the five best passages

Inbrev · 8B

16/16

clinical test messages go to a person, not into automation

Trust, but check: every answer shows its evidence. CheeseDrill, on the charter's promises: 1 December 2026.

WHAT WE LEARNED ABOUT APERTUS

Three lessons for everyone building on it.

1

Openness makes measuring possible.

Apertus publishes its training data catalogue, so we could check which law version its corpus holds. 70B is right on 64 % with the current wording there, 11 % with only the old. Association, not proof.

2

Give it the page, not the book.

8B scores 0.959 with the right passages and 0.720 with the whole booklet. Under a ninth of the tokens, better answers.

3

Confident is not correct.

Label probability is above 0.9 in 94–98 % of answers, and calibrating it gained little. Quotes, numbers and dates get checked in code.

1: Inkraft, 564 and 84 rows by corpus content · 2: GeaNaVid, slide 5 · 3: GeaNaVid development subset, 450 claims.

WHAT'S NEXT

Inbrev becomes a product. The rest stays open.

Product · Inbrev

Nov–Dec 2026	Interviews with 5–10 practice assistants and practice owners: real volumes, fax share
Q1 2027	Pilot in one group practice, on its own hardware
Q2 2027	Written intended purpose and classification review, before the first sale

Open core: the code stays open. Practices pay for operation, connectors and support.

Open source

16 Oct 2026	Inkraft dataset on Hugging Face. GeaNaVid and Inbrev code public
After 1 Dec 2026	CheeseDrill probes and findings, after the Apertus team has seen them
Every release	CheeseDrill and Inkraft as regression tests for the next Apertus

OUR ASK · FOUR TOOLS, THREE PARTNERS

Pilot it. Run it. Break it.

A Swiss group practice: pilot Inbrev on its own hardware in Q1 2027.



The Apertus team: run CheeseDrill and Inkraft on every new release, and tell us which hole to drill next.

OST and voting-information teams: try GeaNaVid on the next booklet and tell us where it is wrong.



Scan to try all four tools

apertool.ch

Tools, video and reports

Niklas Wichter · hello@apertool.ch

Apertus is developed by EPFL, ETH Zurich and CSCS. Hack Apertus is run by the aiLights Association. Apertool is an independent participant.