Inkraft

A trilingual benchmark on Swiss social and health insurance law

Niklas Wichter
Team Apertool · Hack Apertus 2026 · Track 1B — Swiss Voices

Three sentence summary: Swiss residents, AHV branch offices, medical practices and social services need exact amounts, deadlines and responsibilities from Swiss social and health insurance law, and a model that states last year's rule is worse than one that says it does not know. Inkraft is a trilingual (German, French, Italian) short-answer benchmark of 326 questions (948 rows) in which every gold answer is backed by a verbatim quote from the authentic text of a federal act as of 1 October 2026, and for 52 amended provisions the previously valid answer is stored verbatim from the earlier Fedlex version, so that wrong answers split into *stale* (earlier law) and other errors. Graded by a panel of 3 open-weight judges, Apertus 1.5 70B answers 50.8 % of rows correctly but only 7.1 % of the rows whose rule changed with a stored earlier answer, where it gives the earlier law in 30.5 %; with the cited paragraph in the prompt it reaches 99.4 % on the same rows; adding the as-of date to the prompt does not change accuracy significantly.

Track: Apertus Readiness - Track 1B (Swiss Voices, Core Task Intelligence)
Event: Online
Team: Apertool — Niklas Wichter
Demo: project page https://inkraft.apertool.ch; Hugging Face dataset with test items, model responses, judgements and metadata: https://huggingface.co/datasets/NWichter/inkraft

---

Use of Apertus

- **Model:** swiss-ai/Apertus-v1.5-8B and swiss-ai/Apertus-v1.5-70B (Project Apertus et al., 2025; Swiss AI Initiative, 2026a)
- **How it is used:** evaluation (examinee). Apertus never drafts or checks items; Apertus 70B is used once as a *comparison* judge to measure self-preference, never for the reported scores, and translates the French and Italian rows of the human judge sample into German as a reading aid.
- **Where it runs:** hosted Hack Apertus endpoint (OpenAI-compatible); any other endpoint via `LLM_BASE_URL`, `LLM_API_KEY`, `LLM_NAME`

Use Case Description

Every person living in Switzerland is insured under the Federal Health Insurance Act (KVG) and the old-age and survivors' insurance (AHV). Questions about deductibles, contribution rates, waiting periods, family allowances and registration deadlines reach municipal AHV branch offices, front-office staff in medical practices, social services and the insured themselves. Mistakes cost money: a missed switching deadline binds a person to an insurer for another year, a wrong waiting period delays an invalidity insurance (IV) claim. Apertus is an open Swiss model that public bodies can run on their own infrastructure (Swiss AI Initiative, 2026a), so the question is not only whether the model knows this law, but whether it knows the law *in force*.

Social insurance law changes on fixed dates. The minimum child allowance under Art. 5 para. 1 FamZG was CHF 200 per month until 2024 and is CHF 215 since the version of 1 January 2025; the deductible for original drugs with cheaper generics (Art. 38a para. 1 KLV) rose from 20 to 40 %. A model trained on older text may give the earlier figure with confidence. Language models have source-specific effective knowledge cutoffs (Cheng et al., 2024), legal hallucinations are frequent (Dahl et al., 2024), and outdated statutory knowledge has been shown for German law (Prior et al., 2026). Inkraft therefore separates three kinds of failure: an answer that matches the rule before an amendment (*stale*), an answer that matches neither the current nor a stored earlier answer (*other*, including near misses), and an honest refusal (*not attempted*).

Dataset Summary

Inkraft covers 19 federal acts and ordinances on health insurance (KVG, KVV, KLV), old-age, invalidity and supplementary benefits (AHVG, AHVV, IVG, IVV, ELG, ELV), accident insurance and the general part (UVG, ATSG), unemployment and income compensation (AVIG, AVIV, EOG), occupational pensions (BVG, BVV 2, BVV 3), family allowances (FamZG) and continued payment of wages (Art. 324a OR). Each item is a single-turn question with a short gold answer, accepted variants, *near-miss* answers (a related but wrong value) and, for amended provisions, the outdated answer. Every item is anchored to one paragraph: Fedlex ELI with version date and the verbatim quote in the language of the item. The as-of date is 1 October 2026 for all items.

Table 1 places Inkraft among related benchmarks. SwitzerlandQA (Project Apertus et al., 2025) covers Swiss civics including insurance, but as machine-translated multiple choice without sources. Swiss-Bench SBP-002 and 003 cover FINMA, data protection and contract law (Uenal, 2026a; 2026b), LEXam covers law exams (Fan et al., 2025), and Prior et al. measure outdated knowledge of German statutes (Prior et al., 2026). To our knowledge no open benchmark combines Swiss social insurance law, gold answers from three authentic language versions, an as-of date with the verbatim earlier rule, credit for abstention, and an exposure check against the published training data of the evaluated model.

Benchmark	Domain, format	Stale	Abst.
SwitzerlandQA	civics incl. insurance; MC	no	no
Swiss-Bench	FINMA, nDSG, OR; mixed	n. r.	n. r.
LEXam	law exams; open + MC	n. r.	n. r.
Prior et al.	German law; temporal QA	yes	n. r.
Inkraft	**social insurance; DE/FR/IT**	**yes**	**yes**

Table 1: Related benchmarks. Stale: separates outdated from wrong answers. Abst.: credits abstention. n. r.: not reported in the abstract (for SwitzerlandQA: Appendix K of the Apertus report).

Release **v1.0.0** (`data/FREEZE.json`) has **326 German core questions, 948 rows**, 311 cores in all three languages. Question types: numeric 114, yes/no 66, procedural 61, temporal 44, calculation 20 (recomputed by the validator from cited inputs), false premise 18, term

3. Registers: professional 169, lay 157. A coverage matrix (domain ×

question type × change status) is in the data card. The temporal bucket is set per *evidence sentence* over all stored versions: 248 cores cite a sentence unchanged since 2024, 78 a sentence that changed in 2024–2026; **52 of them are STALE pairs** (154 rows) whose earlier answer occurs verbatim in an earlier Fedlex version and differs from the current one. Drafting models: GPT-5.5 via Codex 98 cores, Claude 97, Gemini 86 (disclosed per row) and the open model Nemotron 3 Ultra 45 (32 of the STALE pairs).

Data Collection Method

Truth comes from the statute, never from a language model. The pipeline runs Fedlex → paragraphs → version diff → drafting → automatic checks → independent second check → human spot check → freeze → evaluation. It fetches 20 acts and ordinances from the Fedlex SPARQL endpoint (EOV yielded no item that passed the checks) in German, French and Italian: the consolidation applicable on 1 October 2026, the one applicable at the end of 2023, every version in between and published future versions (`src/fetch_fedlex.py`, URLs and SHA-256 in `data/sources/fedlex_manifest.v2.json`). Articles are split into paragraphs; `src/find_changes.py` diffs every paragraph across versions.

Drafting. 303 cores were drafted per paragraph by Claude, GPT-5.5 (Codex) and Gemini in early October; the gold answer must occur verbatim (or as a normalised number) in the paragraph. Because the version diff was the scarce resource, 160 further temporal candidates were drafted by the open model Nemotron 3 Ultra from ranked version diffs (prompt `nemotron-temporal-v2`, hash stored per row), including four found by a numeric diff of acts outside the first manifest (FamZG, EOV, KLV; the EOV candidates did not pass the checks) and checked against the stored Fedlex versions. French and Italian questions are translations, but their gold answers and evidence come from the authentic French and Italian text of the same paragraph, because all three versions are equally binding (Swiss Confederation, 2023).

Automatic verification (`src/validate_items.py`, every row): schema; the evidence occurs verbatim in the cited paragraph of the version valid on the as-of date; the answer occurs in the evidence; the outdated answer occurs verbatim in an earlier stored version and differs from the current value; a near-miss value valid in an earlier version is rejected; calculation items are recomputed from inputs anchored in the cited text; duplicates; language and value consistency across DE/FR/IT; Swiss spelling and a PII scan. **v1.0.0: 0 errors on 948 rows**, 216 warnings (87 a published later version, 44 threshold words, 23 discretionary rules, 62 other; none blocks a row).

Release rules (`src/auto_review.py`, fixed before any model output was scored; decisions logged per core in `data/review/auto/`): existing cores lose paragraph clusters and near-duplicates (19 dropped) and gain proposed answer variants and outdated answers only when the validator passes (21 patched); a candidate needs no validator error, no duplicate and no "reject" in the development pre-screen, and its correction notes are applied by Nemotron and re-validated (`src/auto_fix.py`). Every remaining core then gets an **independent second check by Qwen3-235B** (`src/second_check.py`): answer supported by the paragraph, unambiguous question, no leak, equivalent languages, valid variants and near misses, and a real (not merely renamed) earlier rule. 15 of 335 cores were flagged; 8 were dropped, 7 calculation items were kept because the checker sees no calculation inputs while the validator recomputes them. Drop rates by drafting model: Codex 4.9 %, Claude 7.6 %, Gemini 9.5 %, Nemotron 71 % (114 of 160 candidates; most changes in the diff pool are renamings such as "ordentliches Rentenalter" → "Referenzalter"). Where the judge panel disagreed on an answer, an open model proposed further answer variants; a variant was added only if its number equals the gold number (or all its words occur in the cited paragraph) and its evidence span is verbatim in the paragraph (`src/panel_fixes.py`, logged with evidence; 7 cores). Ambiguity flags were listed, never applied automatically.

Human check (`src/draw_human_sample.py`, seeds 20261013 for cores and 20261016 for judgements, `src/import_human.py`). A stratified random sample of 35 cores (15 STALE pairs, 20 others) was checked by the curator against the linked paragraph, blind to model outputs: 1 rejected and removed before the freeze (STALE stratum 1 of 15, Wilson 95 % 1–30 %; other stratum 0 of 20, 0–16 %): nmv-0012 (Referenzalter nach Artikel 21 Absatz 1). All other rows are `auto_verified`; rows a human checked are `reviewed`. No item was selected or removed on model outcomes; answer variants were added in 7 cores where the judges disagreed (logged with evidence).

Exposure check. Apertus 1.5 was trained on the Swiss legislation subset `switzerland_lexfind` of MultiLegalPile (Niklaus et al., 2024; Swiss AI Initiative, 2026b). For every row we search the verbatim current evidence and, for amended provisions, the earlier evidence in the German, French and Italian files at the published revision of 12 January 2024; we also count 13-gram overlaps between questions and the corpus (Brown et al., 2020; Sainz et al., 2023). The current wording occurs in 564 rows, only the earlier wording in 81, neither in 277 (26 evidences too short). For the 132 `temporal` rows the corpus holds the current wording for 1 and the earlier one for 73. 55 ques-

tions share a 13-gram with the corpus; they reuse statutory phrasing, the questions themselves are new.

Consent. No personal data is collected; vignettes use roles only. A consent form template for optional reviewers is provided (`docs/consent_form_template.md`).

Dataset Details

The statute texts are not protected by copyright (Art. 5 para. 1 URG) (Swiss Confederation, 2025); Fedlex linked data is offered for free use (opendata.swiss, 2026). Questions, answers and metadata are released under CDLA-Permissive-2.0; the licence is stored per item. The intermediate MultiLegalPile files (CC BY-NC-SA) are analysed locally and not redistributed. The data card follows the Hack Apertus template, Datasheets and Data Cards (Gebru et al., 2021; Pushkarna et al., 2022); a canary GUID in every metadata record and on the dataset card lets model builders exclude the set from training (Srivastava et al., 2023).

Evaluation

Method

Grading follows SimpleQA (Wei et al., 2024): each answer is CORRECT, INCORRECT or NOT_ATTEMPTED; INCORRECT answers carry the subtype STALE (matches the outdated answer), NEAR_MISS or OTHER. Metrics: correct, incorrect and not-attempted rates, correct given attempted (CGA), F-score, hallucination rate I/(C+I), share of stale errors, penalised scores $C - \lambda \cdot I$ ($\lambda = 1, 3$) (Kalai et al., 2025; Wen et al., 2025), language consistency (same grade in DE, FR and IT) (Qi et al., 2023) and the open-book gain, which separates a knowledge gap from a reading gap. Intervals are 95 % Wilson or cluster bootstrap over cores (1,000 resamples); paired comparisons use exact McNemar tests.

Grading has three stages. (1) A deterministic pre-grade extracts numbers, number words, units and refusals in all three languages; it is a hint for the judges, never a grade. (2) A **panel of open-weight LLM judges** grades every main-run answer against gold, variants, outdated and near-miss answers with the same prompt (temperature 0) (Panickssery et al., 2024; Zheng et al., 2023): Qwen3-235B-A22B-2507 (Alibaba, open weights), GLM-5.3 (Z.ai, open weights), DeepSeek V3.2 (DeepSeek, open weights). None of them drafted a released item or is an examinee. The majority grade counts; without a majority the first judge decides (4 of 3792 answers). On 61 constructed answers with known grades every judge passes the 0.95 gate on the three grades (error subtype 57 to 60 of 61; Qwen3-235B-A22B-2507 61/61, GLM-5.3 61/61, DeepSeek V3.2 61/61); Mistral Large 3, planned as the European judge, was rate-limited upstream on OpenRouter throughout and was replaced by DeepSeek V3.2. Inter-judge agreement: Fleiss' $\kappa = 0.95$ (three grades), 0.89 with subtypes; unanimous on 97 % of answers; agreement of each judge with the majority: Qwen3-235B-A22B-2507 99.4 %; GLM-5.3 98.4 %; DeepSeek V3.2 98.7 %. The German ablation rows are graded by Qwen alone, and the paired date and prompt comparisons use Qwen on both sides. (3) A human graded a stratified random 22 % sample of the judgements of the submitted run (Apertus 70B, `closed_dated`: 211 answers) blind to the grades; French and Italian answers were read with a German machine translation (Apertus 70B) next to the original. Agreement with the first judge (on which the sample was drawn): $\kappa = 0.91$ (95 % CI 0.85–0.96, raw 95.3 %; DE 0.89, FR/IT 0.92); with the panel majority $\kappa = 0.92$ (raw 95.7 %) (Bavaresco et al., 2025; Cohen, 1960). As a diagnostic Apertus 70B also graded the `closed_dated` runs: it agrees with Qwen (κ 0.86 on 70B answers, 0.93 on 8B answers) but marks 48 more of its own answers correct, against 19 on the 8B's answers, consistent with self-preference; it is never used for scores.

Evaluation setup and inference parameters

Conditions: `closed_dated` (main): system prompt on Swiss federal law with the date and the instruction to answer briefly and say when unsure; `open_book`: the same plus the cited paragraph; on all German rows also `closed_swiss` (no date) and `closed_plain` (bare question). Parameters: temperature 0, top-p 1, max. 256 output tokens, seed 20261001, reasoning mode off, native chat template of the hosted Hack Apertus endpoint; answers of 2026-10-13, 0 failures after retries. Paid model use (OpenRouter, as reported per call): judges 2.34 USD, second check 0.06 USD; Nemotron was used free of charge.

Results

`closed_dated`	n	8B	70B
All rows	948	33.1 (30–36)	50.8 (48–54)
Changed sentence: correct	232	10.3 (7–15)	16.8 (13–22)
STALE pairs: correct	154	5.2 (3–10)	7.1 (4–12)
STALE pairs: **earlier law**	154	14.3 (10–21)	30.5 (24–38)
STALE pairs, open book	154	96.8 (93–99)	99.4 (96–100)
Corpus has current text	564	41.1 (37–45)	63.6 (60–68)
Corpus has no wording	277	25.6 (21–31)	38.3 (33–44)
Corpus has only earlier	81	7.4 (3–15)	11.1 (6–20)
… of which earlier law	81	24.7 (17–35)	38.3 (28–49)

Table 2: Main results on v1.0.0, % correct unless stated (Wilson 95 % CI), majority of the judge panel (`report/analysis.json`). Corpus rows: exposure bucket of the cited wording in MultiLegalPile. Intervals treat rows as independent (the three languages of a core are correlated; cluster bootstrap by core in `data/results.json`). Incorrect / not attempted overall: 8B 55.4 / 11.5 %, 70B 48.0 / 1.2 %; hallucination rate I/(C+I) 8B 62.6 %, 70B 48.6 % (correct given attempted 37.4 % and 51.4 %); F-score 0.35 and 0.51.

(1) Apertus knows the old law, not the new one. On the 154 STALE-pair rows the 70B is correct in 7.1 % and gives the earlier rule in 30.5 %; for 23 of the 52 amended provisions it states the earlier value in at least one language. The 8B answers less (24 % not attempted) and is stale in 14.3 %. On all 232 rows whose cited sentence changed since 2024 the 70B is correct in 16.8 % and the 8B in 10.3 %. **(2) It is a knowledge gap, not a reading gap.** With the cited paragraph in the prompt the 8B and the 70B answer 95.2 % and 96.7 % of all rows and 96.8 % and 99.4 % of the STALE-pair rows correctly (open-book gain +46 points for the 70B, +62 for the 8B); this also shows that the gold answers are recoverable from the cited text. **(3) Exposure tracks accuracy.** Where only the earlier wording occurs in the published training corpus, the 70B is correct in 11.1 % and stale in 38.3 %; where the current wording occurs, it is correct in 63.6 % (association; bucket, act and change date are confounded). **(4) The date does not help significantly.** On the German rows (first judge on both sides) the

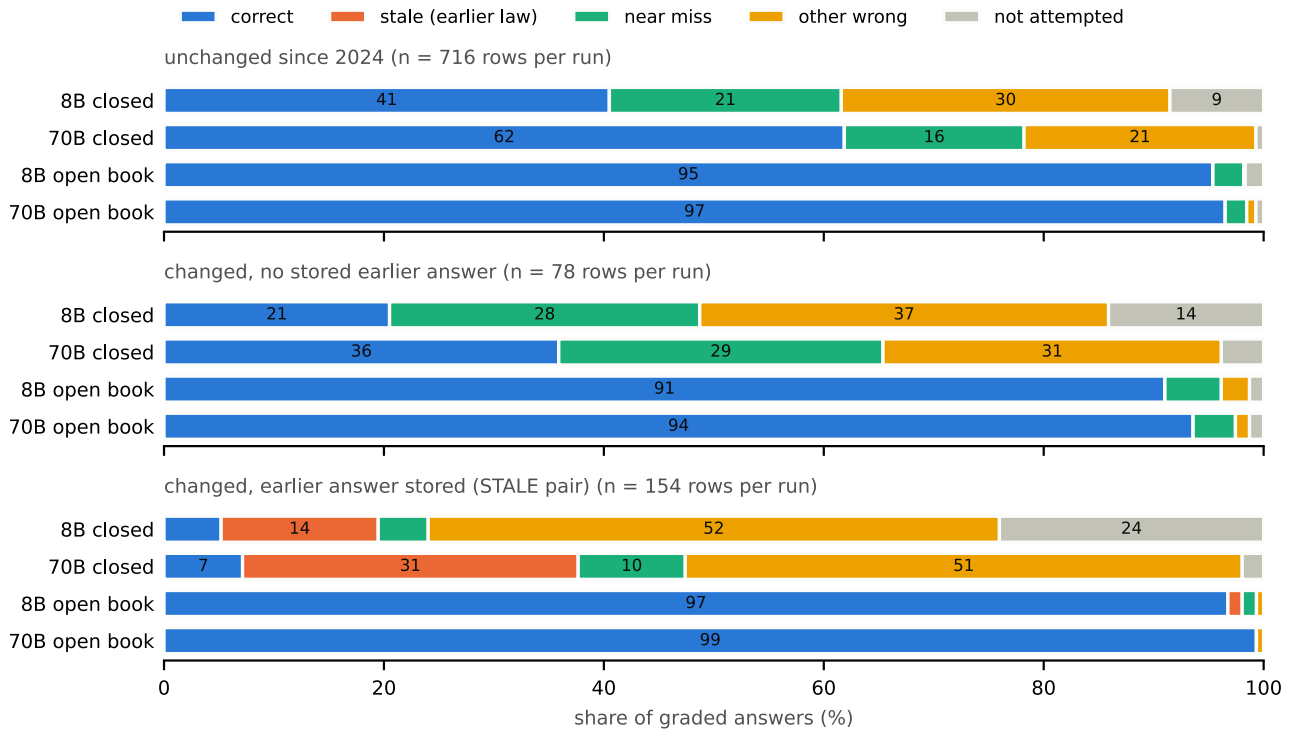


Figure 1: Grades per model and condition for unchanged rules, changed rules without stored earlier answer, and STALE pairs. Numbers are % of rows.

as-of date in the system prompt changes the correct rate by −0.3 points for the 70B (11 vs 12 discordant rows, p = 1.0) and by −1.5 for the 8B (p = 0.27); the full system prompt against the bare question changes it by +1.5 and +1.8 points (p = 0.64, 0.54). **(5) Size helps mostly on stable law.** The 70B is 21 points ahead on unchanged rules (61.9 vs 40.5 %) but 6 on changed ones; overall it beats the 8B on 222 rows and loses on 54 (exact McNemar p < 0.001); by language the 70B is correct in DE 53 %, FR 50 % and IT 49 %; by question type the 70B is correct on 67 % of yes/no items (8B 55 %, guessing rate 50 %), 62 % of calculations, 50 % of false premises and 8 % of temporal questions (table in the data card); it gives the same grade in all three languages for 68 % of the cores (8B: 50 %).

Dataset Limitations

Only federal law; cantonal rules are excluded. Items are verified by deterministic checks against Fedlex, an independent model check and a human spot check of 35 cores, not by a full human review; the second check flagged little (15 of 335), so it is a weak filter for subtle ambiguity. 281 of 326 cores were drafted by closed models (Claude, GPT-5.5, Gemini), disclosed per row; accuracy by drafting model is reported in `data/results.json`, but drafting model and question type are confounded (all calculation and false-premise items come from Gemini, most STALE pairs from Nemotron). French and Italian questions are model translations; answers come from the authentic texts; the French and Italian part of the human judge check relied on a machine translation. One annotator, not a native Swiss speaker. 21 % yes/no items (50 % guessing rate, reported separately). The STALE-pair set is limited by the version diff: most changes since 2024 are renamings or cross-reference renumbering, not new values. The as-of date ages; 29 cores already have a published later version. The exposure check covers only the Swiss legislation subset of MultiLegalPile, not web crawls (Penedo et al., 2025; Su et al., 2025). Short answers measure knowledge, not the quality of advice.

Lessons Learnt and Recommendations

For the model: the as-of date in the system prompt does not make Apertus measurably more current, while the cited paragraph fixes almost every error, so the fix has to come from data or retrieval: a refreshed legislation corpus, retrieval with a version filter instead of parametric recall (Prior et al., 2026), and training abstention instead of rewarding guesses (Kalai et al., 2025). For the dataset: yearly updates on 1 January from the version diff, native-speaker review of French and Italian, a cantonal extension. For benchmark builders: the scarce resource is not drafting but real value changes; a paragraph diff over all intermediate versions plus a verbatim check of the earlier answer found 52 usable STALE pairs, a panel of judges from different model families makes grading less dependent on one model, and Apertus 70B as a judge was more lenient than Qwen overall and more so on its own answers (+48 correct on 70B answers, +19 on 8B answers).

Reproducibility

`make run` (repository root) builds a Docker image (Python 3.12, no GPU, no API keys, a few minutes) and replays the scoring offline: it recomputes every result block from the stored responses (`data/responses/`) and the panel's grades (`data/judgements/`) and compares them with `data/expected_results.json` (PASS/FAIL). `make test` runs the unit tests in the same image.

`make run MODE=full` evaluates any OpenAI-compatible model given `LLM_BASE_URL`, `LLM_API_KEY`, `LLM_NAME` and grades with `JUDGE_*` (smoke test: `LIMIT=50`). `scripts/run_after_freeze.sh` reruns calibration, generation and judging; `src/panel.py` combines the judges; `report/analysis.py`, `report/make_figures.py` and `report/fill_numbers.py` recompute every number and the figure in this report. The dataset construction is scripted end to end (`src/fetch_fedlex.py` → `find_changes.py` → `gen_nemotron.py` → `validate_items.py` → `auto_fix.py` → `auto_review.py` → `second_check.py` → `draw_human_sample.py` / `import_human.py` → `after_review.py` → `panel_fixes.py` → `freeze.py`). Seeds: 20261001 (generation), 20261016 (bootstrap and judgement sample), 20261013 (core sample).

License

This report and the documentation: CC-BY-4.0. Code: Apache-2.0. Dataset (questions, answers, metadata): CDLA-Permissive-2.0, licence stored per item in `metadata.jsonl`. Statute quotes: not protected by copyright (Art. 5 URG). The judge prompt adapts the SimpleQA grader of OpenAI simple-evals (MIT).

References

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., & Testoni, A. (2025). LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 238–255. https://doi.org/10.18653/v1/2025.acl-short.20

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 33 (Neurips 2020)*, 1877–1901. https://arxiv.org/abs/2005.14165

Cheng, J., Marone, M., Weller, O., Lawrie, D., Khashabi, D., & Van Durme, B. (2024). *Dated Data: Tracing Knowledge Cutoffs in Large Language Models*. https://doi.org/10.48550/arXiv.2403.12958

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

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, *16*(1), 64–93. https://doi.org/10.1093/jla/laae003

Fan, Y., Ni, J., Merane, J., Tian, Y., Hermstrüwer, Y., Huang, Y., Akhtar, M., Salimbeni, E., Geering, F., Dreyer, O., Brunner, D., Leippold, M., Sachan, M., Stremitzer, A., Engel, C., Ash, E., & Niklaus, J. (2025). *LEXam: Benchmarking Legal Reasoning on 340 Law Exams*. https://doi.org/10.48550/arXiv.2505.12864

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for Datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). *Why Language Models Hallucinate*. https://arxiv.org/abs/2509.04664

Niklaus, J., Matoshi, V., Stürmer, M., Chalkidis, I., & Ho, D. E. (2024). MultiLegalPile: A 689GB Multilingual Legal Corpus. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15077–15094. https://doi.org/10.18653/v1/2024.acl-long.805

opendata.swiss. (2026). *Nutzungsbedingungen (Terms of Use)*. https://opendata.swiss/de/terms-of-use

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

Penedo, G., Kydlíček, H., Sabolčec, V., Messmer, B., Foroutan, N., Kargaran, A. H., Raffel, C., Jaggi, M., Von Werra, L., & Wolf, T. (2025). *FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language*. https://doi.org/10.48550/arXiv.2506.20920

Prior, M., Schultz, A., & Grabmair, M. (2026). *Asking For An Old Friend: Diagnosing and Mitigating Temporal Failure Modes in LLM-based Statutory Question Answering*. https://doi.org/10.48550/arXiv.2605.23497

Project Apertus, Hernández-Cano, A., Hägele, A., Huang, A. H., Romanou, A., Solergibert, A.-J., Pasztor, B., Messmer, B., Garbaya, D., Ďurech, E. F., Hakimi, I., García Giraldo, J., Ismayilzada, M., Foroutan, N., Moalla, S., Chen, T., Sabolčec, V., Xu, Y., Aerni, M., … Schlag, I. (2025). *Apertus: Democratizing Open and Compliant LLMs for Global Language Environments*. https://doi.org/10.48550/arXiv.2509.14233

Pushkarna, M., Zaldivar, A., & Kjartansson, O. (2022). Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, And Transparency (Facct '22)*, 1776–1826. https://doi.org/10.1145/3531146.3533231

Qi, J., Fernández, R., & Bisazza, A. (2023). Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10650–10666. https://doi.org/10.18653/v1/2023.emnlp-main.658

Sainz, O., Campos, J., García-Ferrero, I., Etxaniz, J., Lacalle, O. L. de, & Agirre, E. (2023). NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for each Benchmark. *Findings of the Association for Compu-*

tational Linguistics: EMNLP 2023, 10776–10787. https://doi.org/10.18653/v1/2023.findings-emnlp.722

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., … Wu, Z. (2023). Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=uyTL5Bvosj

Su, D., Kong, K., Lin, Y., Jennings, J., Norick, B., Kliegl, M., Patwary, M., Shoeybi, M., & Catanzaro, B. (2025). Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2459–2475. https://doi.org/10.18653/v1/2025.acl-long.123

Swiss AI Initiative. (2026b). *Apertus Data Transparency: Versioned Archives of Key Legal Documentation from the Apertus Releases*. https://github.com/swiss-ai/apertus-data-transparency

Swiss AI Initiative. (2026a). *Apertus-v1.5-8B and Apertus-v1.5-70B: Model Cards*. https://huggingface.co/swiss-ai/Apertus-v1.5-8B

Swiss Confederation. (2023). *Bundesgesetz über die Sammlungen des Bundesrechts und das Bundesblatt (Publikationsgesetz, PublG) vom 18. Juni 2004, SR 170.512, Art. 14 Sprachen der veröffentlichten Texte*. https://fedlex.data.admin.ch/eli/cc/2004/745

Swiss Confederation. (2025). *Bundesgesetz über das Urheberrecht und verwandte Schutzrechte (Urheberrechtsgesetz, URG) vom 9. Oktober 1992, SR 231.1, Art. 5 Nicht geschützte Werke*. https://fedlex.data.admin.ch/eli/cc/1993/1798_1798_1798

Uenal, F. (2026b). *Swiss-Bench 003: Evaluating LLM Reliability and Adversarial Security for Swiss Regulatory Contexts*. https://doi.org/10.48550/arXiv.2604.05872

Uenal, F. (2026a). *Swiss-Bench SBP-002: A Frontier Model Comparison on Swiss Legal and Regulatory Tasks*. https://doi.org/10.48550/arXiv.2603.23646

Wei, J., Karina, N., Chung, H. W., Jiao, Y. J., Papay, S., Glaese, A., Schulman, J., & Fedus, W. (2024). *Measuring Short-Form Factuality in Large Language Models*. https://doi.org/10.48550/arXiv.2411.04368

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

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems 36 (Neurips 2023), Datasets and Benchmarks Track*, 46595–46623. https://doi.org/10.52202/075280-2020