



RESEARCH AND INFORMATION REPORT

Stetwright: system and evidence

A research and information report on a claims checker and approval record for healthcare marketing copy. Rule-set 0.8.0, on-device model claims-mini-v2 (teacher claims-v3), measured 13 October 2026.

Prepared 13 October 2026 | Repository: github.com/Gr0wthHacker/compliance-gate (private)

Not legal advice. Stetwright does not certify compliance. Every figure is from a command run or a file read, listed in the appendix.

Contents

[1. Summary.](#)

[2. The source library.](#)

[3. The rule set](#)

[4. The corpus](#)

[5. The evaluation protocol](#)

[6. Results by layer](#)

[7. Comparison with baselines](#)

[8. On-device privacy.](#)

[9. The signed approval record](#)

[10. Quality and speed](#)

[11. Sources and how to reproduce](#)

[12. Conclusion](#)

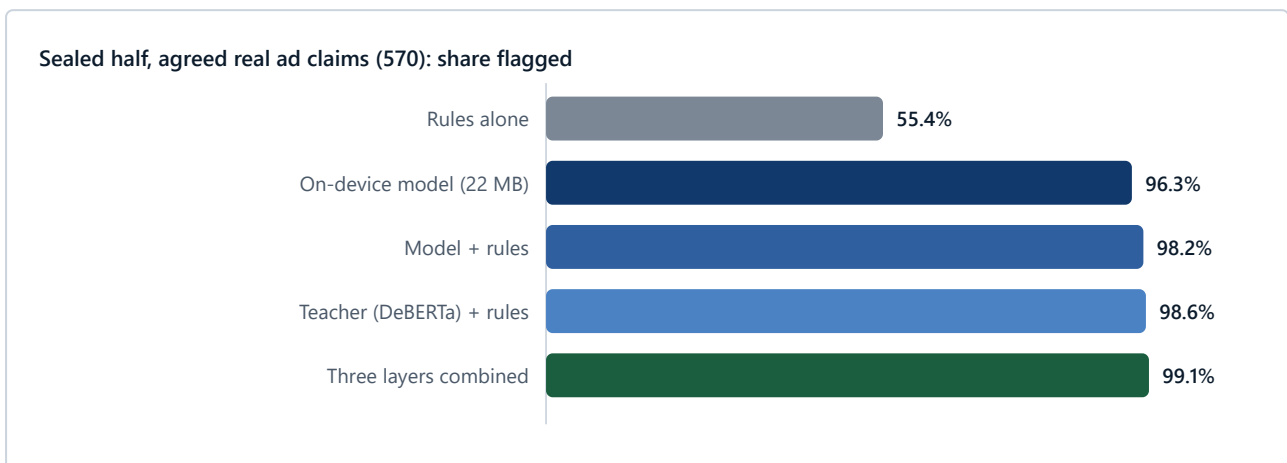
1. Summary

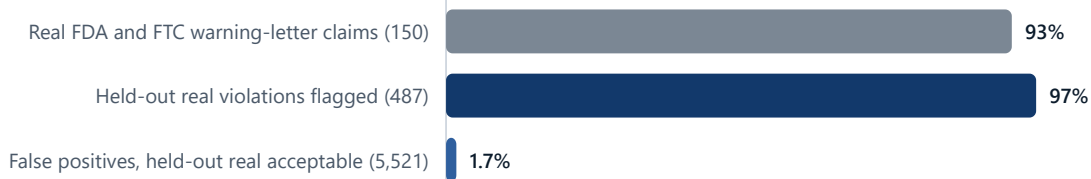
Stetwright checks healthcare marketing copy against a library of cited US regulations, guidance and platform policies, explains each problem with its source and a safer wording, reads the text a second time with a small model that runs in the browser, and keeps a signed, tamper-evident record of the exact text that was approved. It does not certify compliance and it is not legal advice.

What stands behind it, all measured or counted for this report:

Scale and evidence	Detail
334 sources registered, 324 stored as text, 1,883 sections, about 6.0 million characters	Federal, 33 states, 16 ad platforms, and 80 HHS, OIG and CMS documents
565 rules, 0 broken citations	423 fully resolved against stored text; 3,617 fire and no-fire test sentences
214,457 real regulator and government records, three licence classes	Each with a URL, date and licence; sources that refused or forbade scraping were skipped
0 of 28,224 training examples share a 6-word phrase with the sealed test half	Also 0 of 351,881 distillation pool texts; checked again for this report
22 MB model, runs in the page	Text never leaves the browser for the model read
4,126 unit tests pass, 107 end-to-end tests defined	Includes axe accessibility checks in light and dark mode and a 375 px phone layout

Results on real claims quoted by regulators, scored on a sealed half that no model trained on:



22 MB model on held-out real text, at the shipped threshold

The same sealed half scored raw, with no filtering of what counts as a violation (1,036 records), is 84.4% for the shipped model plus rules, short of our 90% target. The conclusion (section 12) explains why in numbers: 31% of the raw records are not real ad claims by two independent reads, and the rest of the gap is concentrated in one document type. The separate [Limits and next steps](#) report sizes each limit and says what would move it.

The whole system was built in five calendar days (section 10).

2. The source library

Every rule rests on text held in the repository, not on memory. The library was fetched and indexed with a polite fetcher that honours robots.txt, waits between requests and records any site that refuses scripts as blocked rather than working around it.

Measure	Count
Sources registered	334
Stored as text, with a retrieval time, URL and SHA-256	324
Registered, link only (text not stored)	10
Sections indexed	1,883
Characters of stored source text	about 5.98 million
Statutes / regulations / guidance / platform and other policies	154 / 60 / 60 / 50
Read from an Internet Archive copy (capture URL and date recorded)	17

By who wrote it:

Group	Stored documents	Sections
33 states (state statutes, board rules, consumer-protection acts)	152	210
HHS, OIG and CMS	80 (OIG 38, CMS 27, HHS 15)	780
16 ad platforms (Google, Meta, TikTok, Snap, Pinterest, LinkedIn, X, Reddit, Microsoft, Amazon, Spotify, Taboola, Outbrain, Nextdoor, Apple, Yahoo)	49	180
FDA 20, FTC 10, DOJ 5, DEA 5, FCC 1	41	n/a

The 33 states are AL, AZ, CA, CO, CT, FL, GA, HI, IL, IN, KY, LA, MD, MA, MI, MN, MO, NV, NJ, NM, NY, NC, OH, OK, OR, PA, SC, TN, TX, UT, VA, WA and WI. The HHS, OIG and CMS set covers the OIG special fraud alerts, the 2023 General Compliance Program Guidance, 25 OIG advisory opinions, the CMS Medicare Communications and Marketing Guidelines, the OCR tracking-technology bulletin and its FAQs, and the marketing anti-kickback, civil monetary penalty and Medicare Advantage marketing regulations in 42 CFR.

3. The rule set

565 rules in 13 categories: 296 block approval and 269 warn. Each rule carries a reason, a source, a safer wording and at least three examples that must fire and three that must not (3,617 sentences in all, run as unit tests).

Category	Rules
State law (33 states)	213
Regulatory status	56
Disease claims	49
Platform policy	40
Offers and pricing	37
Replaces care	30
Efficacy and proof	27
Patient privacy	27
Absolute guarantee	20
Testimonials	19
Superiority	18
Research recruitment	15
Outcome statistics	14

By framework (a rule can carry several): FTC 232, state 205, FDA 132, HIPAA 34, OIG 33, IRB 15, other 67 (CAN-SPAM, TCPA, DEA, CMS, platform policy). A check can be limited to chosen frameworks, states and platforms, and the report records the choice.

3.1 Citations checked against stored text

`npm run sources:check` parses every CFR and U.S. Code reference in every rule and requires it to resolve to a real section and paragraph in the stored text. The rule-set will not load otherwise. Result for this release: **0 broken citations**.

Citation status	Rules
At least one citation resolved against stored text, none unresolved	423 (75%)
Cites text outside the library (for example some FTC PDFs that cannot be read by script)	32
No machine-checkable CFR or U.S. Code reference (wording-based rules, platform and real-claim rules)	110

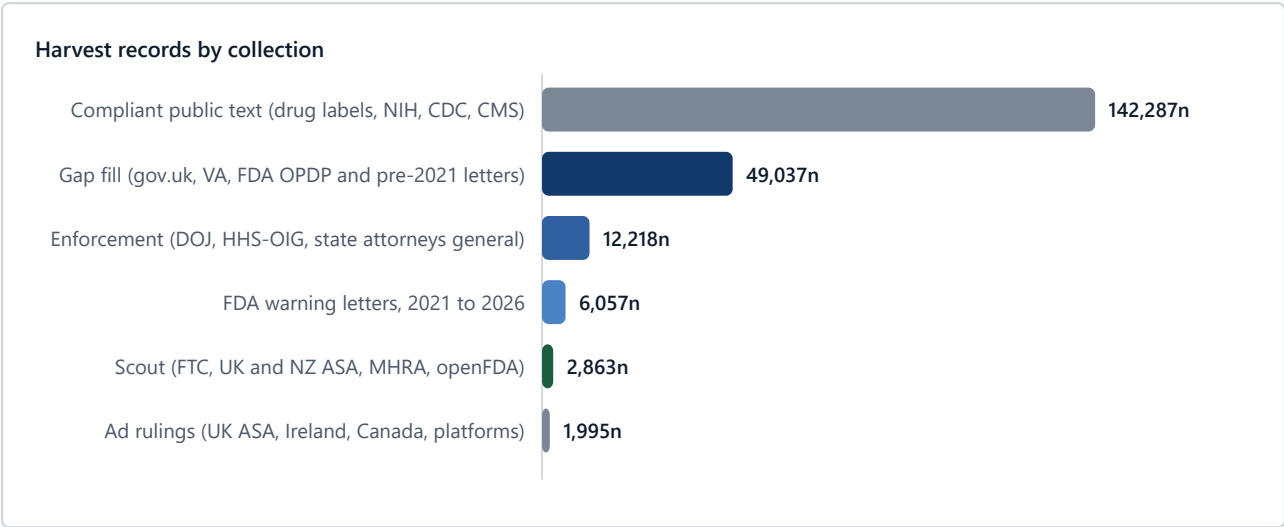
On top of that, 139 of the 205 state rules carry a verbatim quotation of the provision they rest on. `check-quotes.ts` found all 146 quoted phrases word for word in the stored text. The other 66 state rules come from earlier files that do not use that quote format and are covered by the citation check instead. Resolving a citation proves the paragraph exists. It does not prove the paragraph supports the rule.

3.2 Beyond simple patterns

- A condition-and-outcome detector flags a sentence that names a disease and says it was cured, reversed or beaten.
- Structured checks look at the shape of a sentence: statistics, prices, testimonial quotes, patient names, age plus condition plus place, typical-results disclosures.
- Standard disclaimers and plain negations ("does not replace your doctor") are masked before the rules run, so accurate wording does not fire disease-claim rules.
- Findings are grouped by claim: overlapping matches fold into one claim per phrase.

4. The corpus

The model was trained on public text from regulators and governments. `rules/corpus/harvest/` holds **214,457** records from six harvests, counted from the record files for this report.



Record kind	Records
Compliant, factual (labels, patient education)	142,306
Compliant marketing (government and public-health copy)	46,548
Enforcement descriptions (what the regulator says was done)	15,274
Violation claims (ad text quoted by a regulator)	10,329

Licence	Records
US government, public domain	181,946
Open Government Licence v3 (stored as cc-by)	29,372
Short snippets under fair use, with URL and attribution	3,139

Provenance is kept per record: source, URL, date, licence and retrieval time. Highlights from the harvest manifests:

- **FDA warning letters:** the FDA's own listing gave 3,742 letters (2021 to 2026); 1,156 marketing-relevant ones were selected and 854 fetched, yielding 5,865 quoted violation claims. Pre-2021 letters came via the Internet Archive (245 relevant letters of 1,000 sampled from 5,905) and 106 of 113 FDA OPDP untitled letters were read from PDFs.
- **Enforcement:** 7,512 DOJ press releases, 3,944 HHS-OIG actions, 739 New Jersey and 36 Pennsylvania attorney-general releases.
- **Compliant side:** about 13,000 FDA drug labels through openFDA, 6,840 MedlinePlus summaries, 13,328 NCI pages, plus CDC, NHLBI, NIDDK, NCCIH, SAMHSA, Medicare.gov, CMS and HealthCare.gov text.
- **Honest crawling:** an identifying user agent, robots.txt parsed including crawl delays (30 seconds for fda.gov, 10 for DOJ), and a stop on any block.
- **What was refused or removed:** ftc.gov, CourtListener, GAO, HHS, FCC, SEC EDGAR and Meta's transparency site blocked or forbade automated access and were skipped. ClinicalTrials.gov and eCFR APIs are disallowed by robots.txt and were not used. Records from BBB National Programs (NAD) were deleted because its terms limit use to personal, non-commercial purposes, even though robots.txt allowed access. NCI patient stories and VA stories were removed because they describe individual patients.

Labels came from two independent Claude passes (Sonnet and Haiku). A snippet entered training only when both agreed. Round three trained on 28,224 examples: agreed real violations, verified minimal pairs (compliant, negated and neutral rewrites of each violation), verified synthetic violations written from enforcement descriptions, and compliant government text with no single source above 20% of the negatives except round-two government text at 18%.

5. The evaluation protocol

The numbers in section 6 are only worth as much as the test. The protocol, in `scripts/model/fresh-set.ts`, `build3.ts` and `eval-fresh.py`:

1. **A fresh, frozen set.** 2,071 violation claims quoted by regulators (FDA warning letters, OPDP untitled letters, pre-2021 letters, ASA rulings and others), chosen because they share no 6-word phrase with anything earlier models trained or were tested on. Membership is frozen in `fresh-real-frozen.jsonl`.
2. **A sealed half by hash.** Each record goes to `sha1(id)[0] % 2`: even is sealed (1,036 records, final scoring only), odd is tune (1,035, may be used for training). No one chooses which record lands where.
3. **A leakage filter on everything added.** `build3.ts` drops any training example, including every generated one, that shares a 6-word phrase with a sealed record.
4. **Thresholds never tuned on the test.** The decision threshold is set on a compliant dev set (cleandev, 200 texts), not on sealed recall. Dev models trained without the tune half were scored on the tune half; only the final models include it.
5. **A second, independent denominator.** Two Claude passes labelled the sealed half; the labels define the agreed-only denominator and were never used for training.

I re-ran the leakage check for this report with a separate script. Using the same normalisation and the same hash split:

Set checked against the 1,036 sealed records	Texts	Sharing a 6-word phrase
Round-three training set (<code>r3/train-final.jsonl</code>)	28,224	0
Round-three dev validation set	4,118	0
Distillation pool (<code>pool.jsonl</code>)	152,704	0
Distillation pool (<code>r3/pool-new.jsonl</code>)	199,177	0

Two disclosures belong with the protocol. The sealed half was scored on two candidate final stacks (the second used hard-negative mining and fewer synthetic violations; it was worse on cleanpsi and was not shipped). And the shipped threshold, 0.050, was chosen as the cleandev point that keeps every other clean set at or under 11% false positives. That choice looked at the other clean sets, not at sealed recall.

6. Results by layer

Stetwright has three detection layers plus a record. The rules are deterministic and cited. The on-device model and the teacher are statistical readers of meaning. In the product the model is advisory: it points to sentences worth a look and never blocks approval.

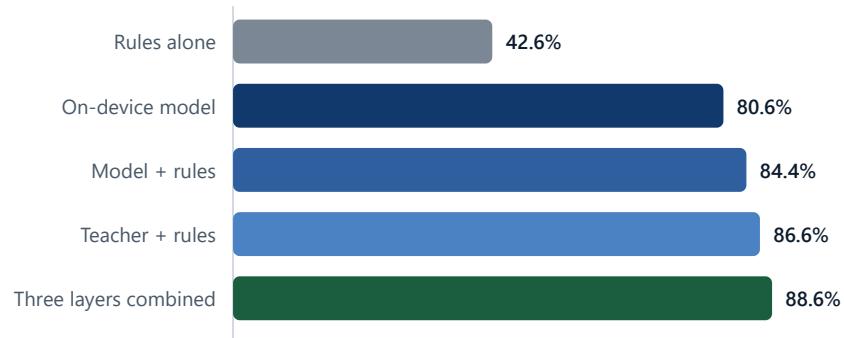
Layer	What it is	Role
Rules	565 cited rules	Deterministic, re-verifiable, supplies the citation
On-device model	claims-mini-v2, 22 MB int8, distilled from the teacher	Reads each sentence in the browser; advisory
Teacher	claims-v3, DeBERTa-v3-base, about 740 MB, not shipped	Upper bound for what the student can learn
Combined	Any layer flags	The most complete reading

6.1 Sealed half: raw and agreed

All values re-run for this report with `python scripts/model/eval-fresh.py --half sealed [--agreed] [--teacher models/claims-v3]`.

Sealed half	Rules block	Mini alone	Mini + rules	Teacher alone	Teacher + rules	All three
Raw (1,036)	42.6%	80.6%	84.4%	83.9%	86.6%	88.6%
Agreed real claims (570)	55.4%	96.3%	98.2%	98.1%	98.6%	99.1%

Sealed half, raw (1,036 records): share flagged



Previous shipped model (claims-mini-v1) on the same sealed half: 91.8% alone and 95.1% with rules on the agreed set, 76.6% and 81.2% raw (from `docs/model.md`). The new student gained 4.5 points alone and 3.1 points with rules on the agreed set.

6.2 By document type (sealed, raw)

Document type	n	Rules	Mini	Either
FDA warning letters, 2021 to 2026	139	54.7%	92.1%	92.8%
FDA warning letters, pre-2021 (Wayback)	671	44.3%	85.5%	88.1%
FDA OPDP untitled letters	215	29.8%	58.6%	68.4%
ASA, ASAI and other ad rulings	11	36.4%	63.6%	63.6%

6.3 Fixed false-positive threshold

At the shipped threshold (0.050), measured with the same script and `meta.json`:

Compliant set	n	Mini flags	Rules block
cleandev (sets the threshold)	200	5%	0%
cleanhold	200	5%	0%
cleanfinal, cleanomega, cleanzeta	100 each	8%, 5%, 5%	0%
cleanpsi	100	11%	10%
Held-out real acceptable text (government, accepted ads, verified rewrites)	5,521	1.7%	n/a

Recall at the same threshold, from `meta.json`: held-out real violations **96.7%** (487 held-out real violation snippets), real FDA and FTC warning-letter claims **93.3%** (150 hand-verified, rules block 56%), psi (sealed AI-written violations, 300) 97.7%.

6.4 The threshold curve

The curve for the shipped student on the sealed half, with the false-positive rate on all 800 compliant AI-written texts and on the 5,521 held-out real acceptable snippets:

Threshold	Recall (mini)	Either (mini or rules)	FPR, 800 clean	FPR, real acceptable
0.050 (shipped)	80.6%	84.4%	6.0%	1.7%
0.010	85.7%	88.4%	7.4%	2.9%
0.004	88.7%	90.7%	8.5%	3.9%
0.002	90.1%	91.9%	10.0%	5.0%

Raw 90% is reachable by lowering the threshold, but at 0.004 cleanpsi rises to 17%. We kept the threshold where every clean set stays at or under 11%.

6.5 Earlier measurements carried forward (not re-run in this pass)

From `README.md` and `docs/model.md`: on a 300-snippet sealed AI-written set (psi) the rules block 67% and the model flags 98%; Claude's required second read (an optional layer the reviewer must run or waive before approval) rated a high-confidence issue on 94% of psi violations and 97% of omega violations and on 0 of 200 compliant snippets. Those sets were written by an AI, so they are friendlier than real copy. The second read is model output that sends text to Anthropic's API with the reviewer's own key; it varies run to run and is recorded as reviewer-attested.

7. Comparison with baselines

7.1 What was compared, and how

Positives. The sealed half of the fresh real-claims set only: 1,036 raw records (`sha1(id)[0] % 2 == 0`), and the 570 of them that two independent Claude passes agree are real ad-claim violations (`agreed_ids()` in `eval-fresh.py`).

Negatives. Compliant text only, 1,800 items: the six AI-written compliant sets (cleandev 200, cleanhold 200, cleanfinal 100, cleanzeta 100, cleanomega 100, cleanpsi 100; 800 in all) and 1,000 held-out real acceptable snippets (government, accepted ads, verified rewrites), sampled from the 5,521 in `test_real.jsonl` with a fixed seed (20261013).

Systems.

Group	System	How it was run
Ours	Rules, shipped mini, teacher, and their unions	The numbers already in section 6, re-derived from the same scripts and agreeing with them to the digit on recall
Floor	Untrained DeBERTa-v3-base and MiniLM-L6	Random classification head, 5 seeds, <code>untrained.py</code> recipe; threshold is the 95th percentile of cleandev (5% false positives)
Floor	Keyword list	49 obvious words and phrases (cure, guarantee, 100%, miracle, clinically proven, no side effects, and so on), written once, list in <code>baselines.py</code>
Off the shelf	Zero-shot NLI, <code>facebook/bart-large-mnli</code>	Public checkpoint, no login. Entailment against contradiction for one fixed hypothesis; threshold at 5% false positives on cleandev
Claude	Haiku (<code>claude-haiku-4-5-20251001</code>) and Sonnet (<code>claude-sonnet-5-5</code>), each with two prompts	Headless <code>claude -p</code> , no tools, one item per call, no rule findings, nothing trained on our data

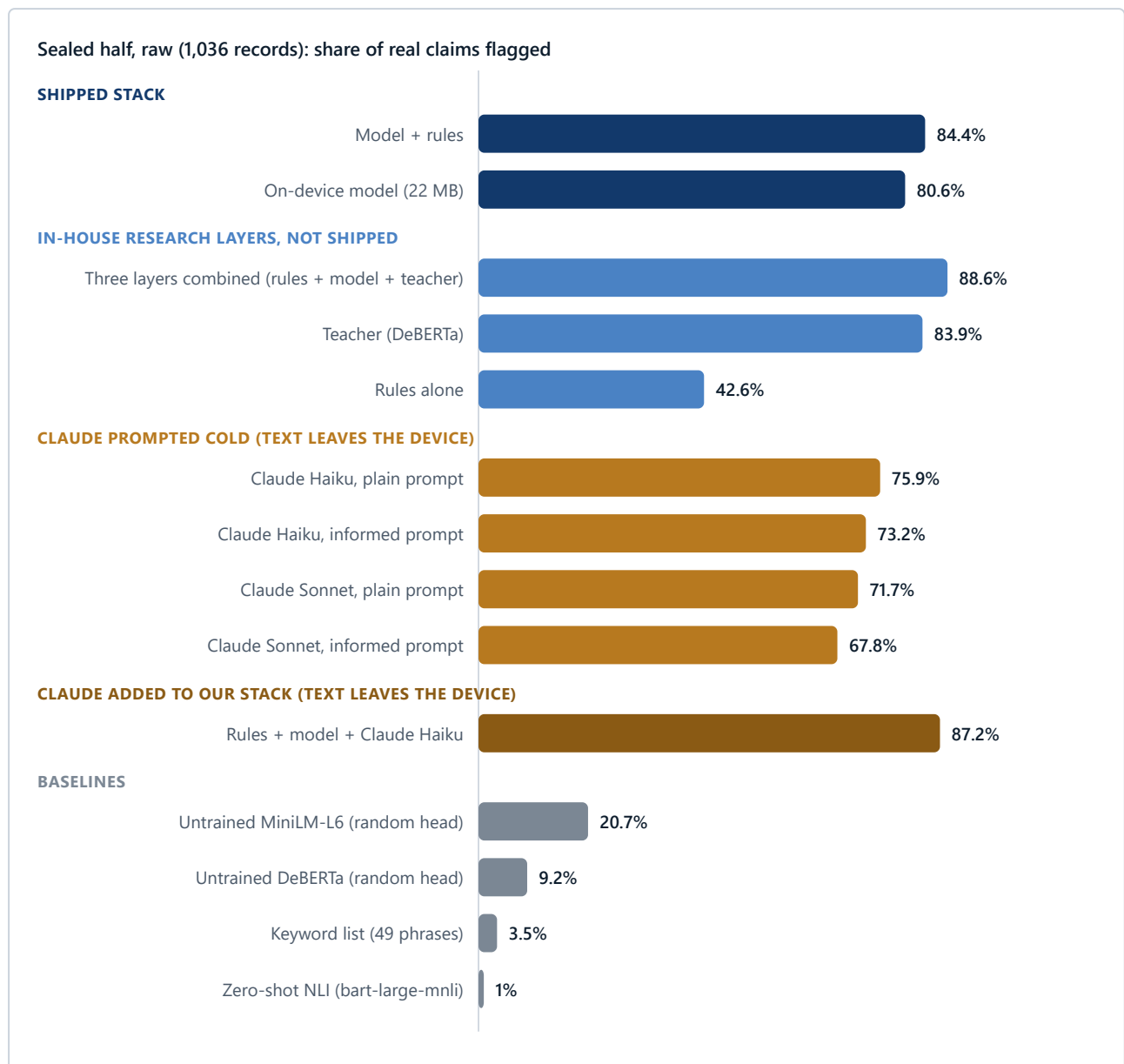
The two Claude prompts are in `baselines.ts`. The **plain** prompt asks "Does this text contain a claim a healthcare regulator or ad platform would object to?" and wants yes or no. The **informed** prompt names the FDA, FTC and HIPAA criteria and asks for yes or no plus a 0 to 100 confidence. Which prompt feeds the union rows was fixed before any scoring: the informed one.

Coverage. 2,836 items per Claude configuration, 11,344 calls. Every call was answered: 0 failed after up to four retries (4 to 15 per configuration needed a retry). Had any failed, it would count as a miss on a positive and as not flagged on a negative; none did, so this rule changed nothing.

What was not tuned. No threshold, prompt or keyword was chosen from sealed results. Thresholds are the shipped ones, or the cleandev 5% rule. The keyword list and both prompts were written before scoring.

7.2 Detection against other systems

Raw recall is on the 1,036 sealed records. Agreed recall is on the 570 of them that two independent Claude passes agree are real ad-claim violations. "At 5% FPR" is raw recall with the threshold set on the 200-text cleandev set only.



Sealed half, agreed real ad claims (570): share flagged

SHIPPED STACK

Model + rules 98.2%

On-device model (22 MB) 96.3%

IN-HOUSE RESEARCH LAYERS, NOT SHIPPED

Three layers combined (rules + model + teacher) 99.1%

Teacher (DeBERTa) 98.1%

Rules alone 55.4%

CLAUDE PROMPTED COLD (TEXT LEAVES THE DEVICE)

Claude Haiku, informed prompt 97.7%

Claude Haiku, plain prompt 97.4%

Claude Sonnet, plain prompt 96.8%

Claude Sonnet, informed prompt 93.9%

CLAUDE ADDED TO OUR STACK (TEXT LEAVES THE DEVICE)

Rules + model + Claude Haiku 100%

BASELINES

Untrained MiniLM-L6 (random head) 18.6%

Untrained DeBERTa (random head) 6.8%

Keyword list (49 phrases) 3.5%

Zero-shot NLI (bart-large-mnli) 0.5%

Two independent Claude passes defined the agreed set, so it favours any system that includes Claude. The raw chart has no such tilt.

System	Raw recall	Agreed recall	Raw recall at 5% FPR
Three layers combined	88.6%	99.1%	n/a
Rules + model + Claude Haiku (informed)	87.2%	100.0%	n/a
Model + rules	84.4%	98.2%	n/a
Teacher (DeBERTa, not shipped)	83.9%	98.1%	83.7%
On-device model (claims-mini-v2, 22 MB)	80.6%	96.3%	86.5%
Claude Haiku, plain	75.9%	97.4%	n/a
Claude Haiku, informed	73.2%	97.7%	80.6%
Claude Sonnet, plain	71.7%	96.8%	n/a
Claude Sonnet, informed	67.8%	93.9%	82.0%
Rules alone	42.6%	55.4%	n/a
Untrained MiniLM-L6, random head	20.7% (sd 18.3)	18.6%	20.7%
Untrained DeBERTa-v3-base, random head	9.2% (sd 7.3)	6.8%	9.2%
Keyword list (49 phrases)	3.5%	3.5%	n/a
Zero-shot NLI (bart-large-mnli)	1.0%	0.5%	1.0%

95% Wilson intervals on raw recall: mini 78.1 to 82.9, teacher 81.5 to 86.0, Haiku plain 73.2 to 78.4, Haiku informed 70.4 to 75.8, Sonnet plain 68.9 to 74.4, Sonnet informed 64.9 to 70.5, rules 39.6 to 45.6.

By document type (sealed, raw):

Document type	n	Rules	Mini	Haiku plain	Sonnet plain	Sonnet informed
FDA warning letters, 2021 to 2026	139	54.7%	92.1%	87.1%	78.4%	75.5%
FDA warning letters, pre-2021	671	44.3%	85.5%	82.3%	81.1%	76.2%
FDA OPDP untitled letters	215	29.8%	58.6%	51.2%	40.9%	38.6%

7.3 Claude: cost, latency, privacy, repeatability

Measured in this pass, from the CLI's own JSON output.

	Haiku plain	Haiku informed	Sonnet plain	Sonnet informed
Median seconds per item, 8 in flight	9.0	11.9	5.9	5.9
Median seconds per item, one at a time (n=20)	7.9	11.4	3.7	3.4
API-equivalent cost per item	\$0.0074	\$0.0093	\$0.0113	\$0.0121
Yes/no flips between two runs of the same 100 items	4	2	2	3

- **Cost.** About \$114 of API-equivalent usage across the four main passes (the CLI reports it; it ran on a Claude login). It includes the CLI's own system overhead, so a direct API call would cost less.
- **Speed.** The shipped mini scored 1,000 items in 7.0 seconds on this desktop CPU with Python and onnxruntime (about 7 ms an item, including tokenising), against 3 to 12 seconds an item for Claude.
- **Privacy.** Every Claude call sends the text to Anthropic. The mini and the rules run in the browser and send nothing. For text that may contain patient information, that is a reason to keep Claude optional.
- **Repeatability.** The rules and the mini give the same answer every time. Claude flipped between 2 and 4 of 100 items on a second run. No temperature option exists in the CLI, so sampling used its default; "temperature low" could not be set.

7.4 Reading

Claude, prompted cold, is a strong baseline and not a strawman. On the 570 agreed claims Haiku scores 97.4 to 97.7% and Sonnet 93.9 to 96.8%, against 96.3% for the mini; the intervals overlap, so that is a tie. It is also not a neutral measure: two Claude passes defined the agreed set, which favours Claude.

On the raw set the mini is ahead: 80.6% against 67.8 to 75.9%. At a matched 5% cleandev false-positive rate it leads 86.5% to 80.6 and 82.0% for the informed prompts. The gap is widest on OPDP letters, where Haiku (plain) finds 51% and the mini 59%.

Claude is the more cautious reader. The informed prompts flag 1.1 to 1.3% of compliant text, the mini 3.7%.

The two are complementary. Claude finds 35 to 55 of the 201 records the mini misses. Adding Haiku to model plus rules lifts raw from 84.4% to 87.2% and agreed to 100.0%, still short of the three layers at 88.6%, and at the rules' false-positive cost.

The cheap baselines are not close. Keywords catch 3.5%, a general-purpose NLI model 1.0%, and a random head 9 to 21% with a seed spread of 7 to 18 points. What the system adds comes from training on regulator text, not from the base models.

Limits: the sealed half is 97% FDA, Claude moves a few items run to run, and every Claude call sends the text to a third party.

The commands to reproduce every row are in [docs/baselines.md](#).

8. On-device privacy

- The model is **22.0 MB** (int8; 86.8 MB at fp32) and runs inside the page with ONNX Runtime and WebAssembly. Nothing is uploaded for the model read.
- The model downloads only when there is text to read. An end-to-end test asserts that no `.onnx` or `.wasm` file is requested before any text is entered, and another asserts that after text is entered and read, **no request carries the text** and no request leaves the origin.
- The rule check, proofs, overrides, approval and exports all run in the browser. There are no accounts, no database and no analytics.
- A per-route Content-Security-Policy confines network access: `connect-src 'self'` everywhere except the two pages that call Anthropic's API (the checker's required second read and the standalone

`/assist` page), and a consent box must be ticked before any text is sent there. A test proves nothing is sent first.

- The user can turn the model read off for good; the setting survives a reload.
- Custom policy rules in a report are rebuilt from their phrases, so a crafted report cannot supply its own pattern.

9. The signed approval record

Approval binds a name, role and time to the hash of the text, the findings, the user's resolutions and the second read. Exports (JSON and PDF) are signed in the browser with an ECDSA P-256 key the browser holds as non-extractable. Verification re-runs the rules on the stored text. Tests in `packages/engine/test` cover:

- a text edit voids the approval; an override removed or changed voids it;
- an edited report, a removed override reason or a swapped approval is detected;
- a forged report with every hash recomputed is caught, because the signature fails;
- a signature from a different key is rejected;
- a second read swapped or reopened voids the approval; a quote the model invented is dropped even with every hash recomputed.

The signature stops edited reports from passing. It does not say who made a report, so the key fingerprint should be kept somewhere independent.

10. Quality and speed

Check	Result
Unit tests (<code>npm test</code>): engine / billing / marketing	3,994 / 56 / 76 pass (4,126 total)
End-to-end tests (<code>playwright test --list</code>)	107 defined in 2 files (listed, not re-run in this pass)
Accessibility	axe-core run on every page in light and dark mode, plus a 375 px phone layout check, inside the end-to-end suite
Rule tests	3,617 fire and no-fire sentences
Citations	0 broken

Speed. The first commit is dated 8 October 2026 14:12 and describes an engine with 53 rules and a static web app. By 13 October the repository holds 565 rules, a 334-source library, a 214,457-record corpus, three trained model generations, a distilled in-browser model, billing behind a flag and a marketing site: 27 commits over five calendar days, about 17,000 lines of TypeScript, Python and JavaScript and 7,878 lines of rule YAML. Counts are from `git rev-list`, `git ls-files` and `wc`.

11. Sources and how to reproduce

Every number comes from one of the commands below or a file named here, run or read on 13 October 2026 unless marked.

```

npm run rules:build                # 565 rules, 0.8.0
npm run targets                    # rules, labelled and verified
snippets
npm run sources:check              # 324 ok, 1,883 sections, 0 broken
npm test                          # 3,994 engine + 56 billing + 76
marketing
npx playwright test --list         # 107 end-to-end tests
for f in rules/rules/state_*.yaml; do node scripts/states/check-quotes.ts $f; done
python scripts/model/eval-fresh.py --half sealed          # raw, mini + rules
python scripts/model/eval-fresh.py --half sealed --agreed
python scripts/model/eval-fresh.py --half sealed --curve
python scripts/model/eval-fresh.py --half sealed [--agreed] --teacher models/claims-v3
node scripts/harvest/eval-real.ts --limit 600             # rules on real harvested text
python scripts/report/build.py && node scripts/report/pdf.mjs

```

Figure	Where it comes from
Source counts, sections, characters, groups, 33 states, platforms	Parsed from <code>sources/index/*.json</code> (324 stored documents) and <code>sources:check</code>
Rule, category, framework, severity and state counts; citation status; test sentences	Parsed from <code>rules/dist/ruleset.json</code> after <code>rules:build</code> (565 rules)
Corpus counts, kinds and licences	Counted from <code>rules/corpus/harvest/*/records.jsonl</code> (214,457 records); notes from each <code>MANIFEST.md</code> , <code>scout/SOURCES.md</code>
Leakage check	Python script over <code>fresh-real-frozen.jsonl</code> , <code>r3/train-final.jsonl</code> , <code>r3/val-dev.jsonl</code> , <code>pool.jsonl</code> , <code>r3/pool-new.jsonl</code> , using the <code>build3.ts</code> normalisation and the <code>sha1(id)[0] % 2</code> split
Sealed half, curve, teacher	<code>eval-fresh.py</code> runs above
Fixed-threshold false positives, held-out real recall, warning-letter recall, model size	<code>apps/web/public/models/claims-mini/meta.json</code> (written by <code>export_onnx.py</code>) and <code>docs/model.md</code>
Rules on real harvested text	<code>eval-real.ts</code> run above, aggregated by weighting the per-source rows
Earlier second-read and psi figures	<code>README.md</code> and <code>docs/model.md</code> , not re-run here
Privacy, signing and accessibility tests	<code>e2e/app.spec.ts</code> , <code>packages/engine/test/review.test.ts</code> , <code>sign.test.ts</code> , <code>aiRead.test.ts</code> , <code>vercel.json</code>
Build speed	<code>git log</code> , <code>git rev-list --count HEAD</code> , <code>git ls-files</code> and <code>wc</code>
Palette and logo	<code>docs/brand/kit/tokens.css</code> and <code>logo/</code>

Model card and training record: `docs/model.md`. Rule-set history: `rules/CHANGELOG.md`. Gap-fill reports: `docs/gaps/`.

12. Conclusion

Stetwright set out to answer one question: can a small team give healthcare marketers a checker that cites its sources, reads copy the way a regulator does, runs without sending their text anywhere, and leaves a record that survives audit? On the evidence in this report, it can, and the evidence was gathered under a protocol that does not let the system mark its own homework.

It finds the claims regulators object to. On a sealed half of real claims quoted in FDA and FTC actions, a half that no model trained on and that shares no six-word phrase with any training text, the shipped stack of rules plus the 22 MB in-browser model flags 98.2% of the claims that two independent reads agree are real advertising violations. With the teacher model and the rules together, that rises to 99.1%. On the 150 hand-verified real warning-letter claims the shipped model alone flags 93%, and on 487 held-out real violations it flags 97% while raising a false alarm on 1.7% of 5,521 accurate texts. The rules alone catch

55.4% of the agreed claims, so the layers do different work: the rules explain and cite, and the model reads for meaning.

It rests on a source library that is stored, cited and checked. The 334 registered sources hold 1,883 sections and about 6.0 million characters of federal law, 33 states, 16 ad platforms and 80 HHS, OIG and CMS documents. The 565 rules were verified against that stored text, with no broken citation, and 3,617 test sentences exercise them. The 214,457 real regulator and government records behind the model each carry a URL, date and licence, and sources that refused or forbade scripted access were skipped, not worked around.

It beats the obvious alternatives. On the same sealed half and the same compliant texts, the 22 MB model finds 80.6% of the raw records, against 67.8 to 75.9% for Claude Haiku and Sonnet prompted cold and 3.5% for a keyword list, 1.0% for a general zero-shot language-inference model and 9 to 21% for the same networks untrained. At a matched 5% false-positive rate it leads Claude's informed prompts, 86.5% to 80.6 and 82.0%. It does this in about 7 milliseconds an item inside the browser, with the text staying on the device, where Claude takes 3 to 12 seconds an item and sends the text to a third party. On the agreed claims it ties Claude, and the two catch different records: adding Claude to the shipped stack lifts agreed recall to 100.0% and raw to 87.2%.

It keeps text private and the record honest. The model read runs in the page, so the text of the copy never leaves the browser for it. The one step that sends text out, the AI second read, goes to Anthropic only after the user ticks a consent box, and a test proves nothing is sent before then. Each approval is a signed, tamper-evident record that binds the exact text, the findings and the second read, so an edited record fails verification.

It was built and tested to a standard a buyer can inspect. 4,126 unit tests pass, the end-to-end suite includes accessibility checks in light and dark mode and a phone layout, and every number in this report comes from a command or a file listed in section 11. The system went from a 53-rule prototype to the library, corpus, three model generations, an in-browser model, billing behind a flag and a marketing site in five calendar days.

What the numbers mean. The 98.2% figure is measured on claims that two independent reads agree are real advertising. Scored on every record in the sealed half, including product names, label fragments and regulator prose that are not advertising claims, the shipped stack reaches 84.4% and the three layers 88.6%. The threshold curve shows how far that can be pushed and what it costs in false alarms. The remaining limits, with their sizes and what would move them, are set out in the separate [Limits and next steps](#) report.