



RESEARCH AND INFORMATION REPORT

Stetwright: system and evidence (short version)

A short research report on a claims checker and approval record for healthcare marketing copy. Rule-set 0.8.0, on-device model claims-mini-v2, measured 13 October 2026. The full report is `research-report.pdf`.

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. What it is](#)

[2. The scale behind it](#)

[3. How it was tested](#)

[4. Results](#)

[5. Comparison with baselines](#)

[6. Privacy and the record](#)

[7. Reproduce](#)

[8. Conclusion](#)

1. What it is

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 22 MB 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.

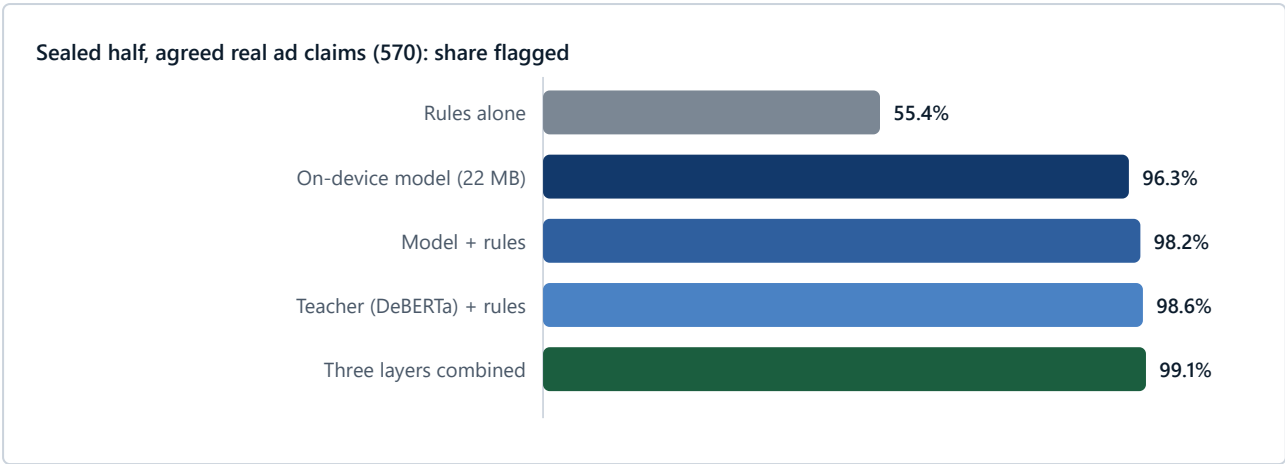
2. The scale behind it

Scale and evidence	Detail
334 sources registered, 324 stored as text, 1,883 sections	Federal, 33 states, 16 ad platforms, 80 HHS, OIG and CMS documents; each stored document has a URL, retrieval time and SHA-256
565 rules (296 block, 269 warn) in 13 categories	0 broken citations; 423 fully resolved against stored text; 3,617 fire and no-fire test sentences
214,457 real regulator and government records	181,946 US public domain, 29,372 Open Government Licence, 3,139 short fair-use snippets; URL, date and licence on every record
4,126 unit tests pass, 107 end-to-end tests defined	Includes axe accessibility checks in light and dark mode and a phone layout
Built in five calendar days	First commit 8 October 2026, 27 commits, 565 rules from a first release of 53

3. How it was tested

- **A frozen, fresh test set:** 2,071 real claims quoted by regulators, none sharing a 6-word phrase with earlier training or test data.
- **A sealed half by hash:** `sha1(id)[0] % 2` puts 1,036 records in a sealed half used only for final scoring.
- **A leakage filter on everything added to training,** including generated text. I re-checked it for this report: 0 of 28,224 training examples, 0 of 4,118 validation texts and 0 of 351,881 distillation texts share a 6-word phrase with a sealed record.
- **Thresholds never tuned on the test:** the decision threshold is set on a compliant dev set of 200 texts, not on sealed recall.

4. Results



Sealed half	Rules	Mini	Mini + rules	Teacher + rules	All three
Agreed real claims (570)	55.4%	96.3%	98.2%	98.6%	99.1%
Raw, nothing filtered (1,036)	42.6%	80.6%	84.4%	86.6%	88.6%

At the shipped threshold the 22 MB model flags 93.3% of 150 hand-verified FDA and FTC warning-letter claims (the rules block 56%) and 96.7% of held-out real violations, while flagging 1.7% of 5,521 held-out real acceptable snippets. On 300 sealed AI-written violations it flags 97.7%. The model is advisory: it points to sentences worth a look and never blocks approval. The rules, with their citations, decide.

5. Comparison with baselines

Same sealed half, same compliant texts, thresholds fixed before scoring (full method and intervals in [baselines](#)).

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

SHIPPED STACK

Model + rules 84.4%

On-device model (22 MB) 80.6%

IN-HOUSE RESEARCH LAYERS, NOT SHIPPED

Three layers combined 88.6%

Rules alone 42.6%

CLAUDE PROMPTED COLD (TEXT LEAVES THE DEVICE)

Claude Haiku, plain prompt 75.9%

Claude Sonnet, informed prompt 67.8%

BASELINES

Untrained MiniLM-L6 (random head) 20.7%

Untrained DeBERTa (random head) 9.2%

Keyword list (49 phrases) 3.5%

Zero-shot NLI (bart-large-mnli) 1%

System	Raw recall (1,036)	Agreed recall (570)	False positives (1,800 compliant)
Three layers combined	88.6%	99.1%	22.1% (no document type declared)
Model + rules	84.4%	98.2%	20.4% (no document type declared)
On-device model, 22 MB	80.6%	96.3%	3.7%
Claude Haiku, plain prompt	75.9%	97.4%	4.9%
Claude Haiku, informed prompt	73.2%	97.7%	1.3%
Claude Sonnet, informed prompt	67.8%	93.9%	1.1%
Rules alone	42.6%	55.4%	17.1% (no document type declared)
Untrained MiniLM-L6 / DeBERTa, random head	20.7% / 9.2%	18.6% / 6.8%	10.3% / 5.4%
Keyword list (49 phrases)	3.5%	3.5%	11.3%
Zero-shot NLI (bart-large-mnli)	1.0%	0.5%	4.9%

The on-device model leads Claude prompted cold on raw recall (80.6% against 67.8 to 75.9%) and at a matched 5% false-positive rate (86.5% against 80.6 and 82.0%), runs in about 7 ms an item in the browser

and keeps the text on the device. Claude takes 3 to 12 seconds an item and sends the text to a third party. On agreed claims they tie, and they catch different records: adding Claude to the shipped stack lifts agreed recall to 100.0%. The cheap baselines are not close, and the untrained networks show the gain comes from training on regulator text.

6. Privacy and the record

- **On-device:** the 22.0 MB model runs in the page. It downloads only when there is text to read, and an end-to-end test asserts that no request carries the text and none leaves the origin.
- **No accounts, no database, no analytics.** A per-route Content-Security-Policy allows `api.anthropic.com` only on the two pages that need it, behind a consent box.
- **Signed approval record:** approval binds a name, role and time to the hash of the text, the findings and the resolutions. Exports are signed with a non-extractable ECDSA P-256 key. Tests show that an edit voids the approval, that edited reports and forged hashes fail verification and that a swapped second read fails.

7. Reproduce

```
npm run rules:build && npm run sources:check && npm test
python scripts/model/eval-fresh.py --half sealed [--agreed] [--curve] [--teacher
models/claims-v3]
node scripts/harvest/eval-real.ts --limit 600
python scripts/report/build.py research-report-short.md research-report-short.html && node
scripts/report/pdf.mjs research-report-short.html research-report-short.pdf
```

8. Conclusion

Stetwright set out to give healthcare marketers a checker that cites its sources, reads copy the way a regulator does, keeps their text in the browser, and leaves a record that survives audit. The evidence supports it. On a sealed half of real claims quoted in FDA and FTC actions, which no model trained on and which 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 two independent reads agree are real advertising violations, and 99.1% with the teacher model added. On 150 hand-verified real warning-letter claims the shipped model flags 93%, and on 487 held-out real violations 97%, with 1.7% false alarms on 5,521 accurate texts. It rests on 334 stored sources, 565 rules with no broken citation, and 214,457 real records with provenance, and it was built and tested (4,126 unit tests, accessibility checks in both themes) in five calendar days.

The 98.2% is measured on claims two independent reads agree are real advertising. Scored on every sealed record, including product names and regulator prose that are not claims, the shipped stack reaches 84.4%; the limits are sized in the separate [Limits and next steps](#) report.