

A claims checker built on real regulator data, tested on text it never saw

Checks healthcare marketing copy against 565 cited rules, reads it with a 22 MB model that runs in the browser (text never leaves), and signs a tamper-evident approval record. Rule-set 0.8.0, model claims-mini-v2. Not legal advice; it does not certify compliance.

98.2%

agreed real claims flagged by the 22 MB model plus rules (570 sealed records)

93%

of 150 hand-verified FDA and FTC warning-letter claims flagged by the 22 MB model

97% / 1.7%

held-out real violations flagged / real acceptable text flagged

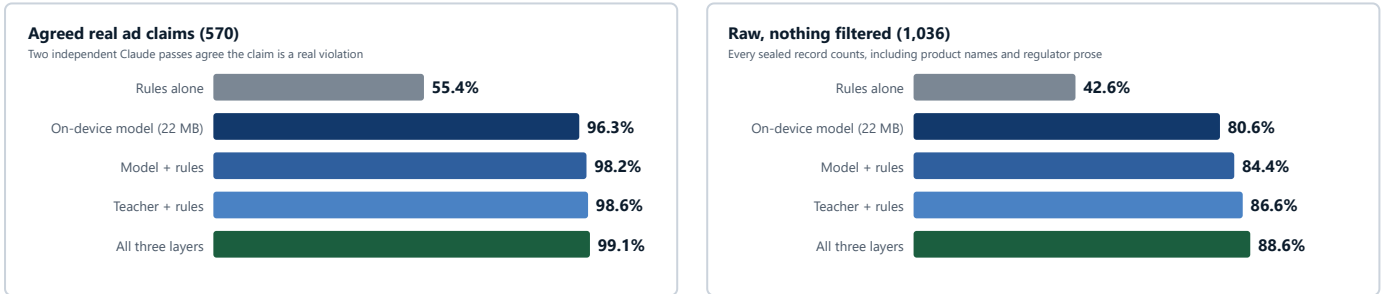
0 of 28,224

training examples share a 6-word phrase with the sealed test half

Scale

What	Count
Sources stored as text (federal, 33 states, 16 ad platforms, 80 HHS, OIG and CMS documents)	324 of 334 registered, 1,883 sections
Rules (296 block, 269 warn) in 13 categories; 0 broken citations, 423 fully resolved against stored text	565
Real regulator and government records with URL, date and licence (US public domain, OGL v3, short fair-use snippets)	214,457
Tests: unit tests passing / end-to-end tests defined (axe accessibility, privacy, signing)	4,126 / 107
Time to build: first commit to this report	5 calendar days

Layered results on a sealed half of real regulator claims



Protocol. 2,071 frozen real claims; sealed half by sha1(id) parity, scored once at the end; every training example checked against a 6-word-phrase filter (re-run: 0 overlaps in 28,224 training and 351,881 distillation texts); thresholds set on a 200-text compliant dev set, never on test recall. The model is advisory and never blocks approval; the cited rules decide.

Against baselines (same sealed half, same compliant texts): on-device model **80.6%** raw, Claude Haiku 75.9% and Sonnet 67.8% prompted cold, keyword list 3.5%, zero-shot NLI 1.0%, untrained networks 9 to 21%. The model runs in about 7 ms an item in the browser; Claude takes 3 to 12 s and sends the text out. On agreed claims they tie (96.3% against 94 to 98%).

Conclusion. On sealed real claims from FDA and FTC actions that no model trained on, rules plus the 22 MB in-browser model flag **98.2%** of the claims two independent reads agree are real advertising violations (99.1% with the teacher model), and 93% of 150 hand-verified warning-letter claims, with 1.7% false alarms on 5,521 accurate texts. It rests on 334 stored sources, 565 cited rules with no broken citation and 214,457 real records, and was built in five days. Scored on every record, including product names and regulator prose that are not claims, the stack reaches 84.4%. Details and limits in docs/research-report.pdf.