



RESEARCH AND INFORMATION REPORT

Stetwright: limits of claim detection

How far the claims checker can go on regulator-quoted healthcare marketing claims, what would push it further (content, models, human labels), and what is a hard limit. Rule-set 0.8.0, claims-mini-v2 and 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

<u>1. Answer</u>	
<u>2. Scoreboard</u>	
<u>3. What the raw misses are</u>	
<u>4. Real upper limits</u>	
<u>5. What more content would add</u>	
<u>6. What more model work would add</u>	
<u>7. Ranked plan</u>	
<u>8. The ceiling, stated plainly</u>	

1. Answer

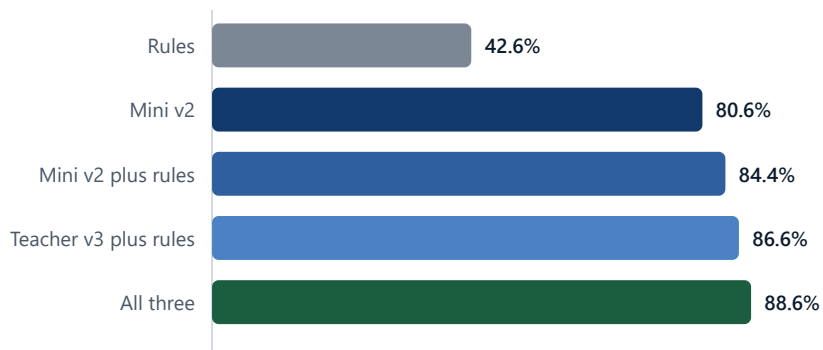
- **Today, on the sealed half (1,036 regulator-quoted claims): 84.4% raw and 98.2% agreed-only** (rules plus the shipped on-device model), at a 7.1% false-positive rate on 800 compliant test texts.
- **Raw is capped by the labels, not by the model.** 322 of 1,036 records (31%) are called a violation by neither of two independent Claude passes (table labels, SOP titles, product numbers, regulator prose). They make up 141 of the 162 raw misses (87%). Only 21 misses (2.0% of all records) are text at least one pass called a violation.
- **Realistic ceiling (estimate): about 88% raw at 7% false positives, 92 to 94% at 10%, 96 to 97% at 15 to 20%. About 99% agreed-only.** Above that, a higher score means flagging text no reader could judge on its own.
- **Which lever?** Models first (a larger student is worth about 2 raw points). Human labels second (they fix the denominator and noisy targets). More public regulator quotes third: that pool is mostly used. Content of a different kind (whole ads, pilot copy, overrides) matters most for the product, and the sealed benchmark will barely register it.
- **Caveat on everything below.** The sealed half is 97% FDA (1,025 of 1,036 records). It measures how well we find claims FDA quoted. It does not measure real small-business copy, UK, state or FTC law, or false positives on real marketing.

2. Scoreboard

Sealed half, thresholds fixed on the compliant dev set, never on sealed recall. Commands: `python scripts/model/eval-fresh.py --half sealed [--agreed] [--teacher models/claims-v3] --curve .`

Layer	Raw (n=1,036)	Agreed-only (n=570)	False-positive rate, 800 AI-written compliant texts
Rules (block)	42.6%	55.4%	1.3%
Mini v2 alone (22 MB, in the browser)	80.6%	96.3%	6.0%
Mini v2 plus rules (shipped)	84.4%	98.2%	7.1%
Teacher v3 alone (DeBERTa, not shipped)	83.9%	98.1%	6.8%
Teacher v3 plus rules	86.6%	98.6%	7.9%
Mini, teacher and rules, any hit	88.6%	99.1%	10.5%

Raw sealed recall by layer (n=1,036)



The 6.0% in `docs/model.md` is the model alone. With rules it is 7.1%, and 10.5% for the three-way union. The headline raw number is at 7.1%. On real government text (18,513 snippets, `node scripts/harvest/eval-real.ts`) rules alone block 20.8%, so rules carry a real false-positive cost that the AI-written clean sets understate.

Per source, sealed (mini plus rules; teacher plus rules; all three; agreed-only mini plus rules):

Source	n	Mini+rules	Teacher+rules	All three	Agreed n and recall
FDA warning letters 2021 to 2026	139	92.8%	95.7%	95.7%	89, 100%
FDA warning letters pre-2021	671	88.1%	89.3%	90.6%	457, 98.5%
FDA OPDP untitled letters	215	68.4%	72.6%	78.1%	23, 87.0%
UK ASA, Irish ASAI	11	63.6%	81.8%	81.8%	1, 100%

Only 11 sealed records are non-FDA. A UK, state or FTC number cannot be read from this half. For scale, the tune half holds 270 UK ASA records, where rules block 34 to 35% (`eval-real.ts`). Those were used in training.

Recall against false positives (model alone unless stated; thresholds swept; false-positive rate on the 800 compliant texts; "real" is 5,521 held-out real acceptable snippets, mostly government text):

FPR budget	Mini raw	Mini+rules raw	Teacher raw	Teacher+rules raw	Mini+rules agreed	Teacher+rules agreed	Real acceptable flagged (mini / teacher)
3%	64.8%	72.2%	72.5%	77.1%	92.6%	96.3%	0.4% / 0.4%
5%	75.9%	80.3%	80.8%	84.0%	97.2%	97.7%	1.3% / 1.1%
10%	90.1%	91.9%	87.5%	90.0%	99.3%	99.3%	5.1% / 2.3%
15%	94.2%	95.5%	91.1%	93.2%	99.6%	99.5%	8.8% / 3.7%
20%	95.5%	96.5%	93.6%	95.5%	99.8%	99.6%	11.7% / 6.2%

The "plus rules" columns add rules to the model threshold for that budget, so their real false-positive rate is about 1.3 points higher.

Reading the table: at a 5% budget the teacher beats the mini by 4.9 raw points (80.8 against 75.9). At 10% the mini is ahead (90.1 against 87.5). The union of all three (88.6% at 10.5%) is no better than simply lowering the mini threshold (mini plus rules: 90.7% at 8.5% model false positives, 91.9% at 10%). Ensembling buys what thresholding buys.

3. What the raw misses are

All 162 raw misses of the shipped stack (mini plus rules) were sorted using the two Claude labels already on file, then the 21 plausible misses and a 45-record random sample of the 141 others were read by hand. **Sealed text was read for this report only. Nothing from it went into training, rules, thresholds or tests, and it will not.**

Bucket	Misses	Share of 162	Share of 1,036
Not a claim by both passes: non-ad text (SOP titles, lot numbers, table headers, section headings)	71	44%	6.9%
Not a claim by both passes: regulator prose quoted back	35	22%	3.4%
Not a claim by both passes: benign ad-like text (product descriptions, instructions, trial-design labels, "ask your provider")	35	22%	3.4%
Violation by both passes, missed	10	6%	1.0%
Violation by one pass, missed	11	7%	1.1%

The 21 plausible misses, by my hand classification:

Type	Count	Example shape
Implied claim, no explicit disease word	9	dosing instructions for a symptom; "prevention is possible"; "may also treat"
Superiority or replaces care	5	"preferred treatment option"; women "choose not to do mammography"
Context-dependent (truncated, or the fault is elsewhere in the ad)	5	a trial statistic with no comparator; a testimonial cut mid-sentence
Regulatory status	1	"approved by the F.D.A." (a rule gap we can close)
Product listing	1	a compound named in a product list

Four findings.

- 1. True model errors are about 2% of records.** The mini plus rules catches 97.1% of the 714 records that at least one pass calls a violation (98.9% with the teacher added). Fixing all 21 would move raw from 84.4% to 86.4%.
- 2. The rest of the raw gap is a definition problem.** The 322 non-claims are quotes from a letter's body, not claims. The models do flag 181 of them (56%), mostly because medical vocabulary and letter-style fragments resemble the training violations. That flatters raw recall. Some of the raw "recall" is recognising the source's style.

3. **OPDP is mostly a labelling story.** Of 215 OPDP records, 154 (72%) are non-claims by both passes and only 23 are agreed violations. Of the 68 OPDP misses, 61 are non-claims and 3 agreed. On the 61 plausible OPDP records the recall is 88.5%. OPDP letters are about risk presentation and fair balance, which is often what the ad leaves out. A snippet that says "Mean number of unscheduled spotting days" is not wrong by itself.
4. **Jurisdiction mismatch is not the cause of these misses.** Only 4 of 162 are UK or Irish. It is a real limit for use, but the sealed half cannot size it.

4. Real upper limits

Hard means no amount of work removes it. **Bounded** means work reduces it.

Limit	Size	Class
Label disagreement. Two Claude passes agree on "violation" for 570 records, 86.1% raw agreement, kappa 0.71. 144 records have one yes. Regulators themselves apply different standards (FDA, FTC, ASA, platforms).	14% of records are disputed; 31% are non-claims	Hard for the benchmark. Bounded for the product.
Snippet-only context. The model reads up to 128 tokens. Fair balance, risk omission and net impression need the whole ad.	5 of 21 plausible misses; many of the 35 benign-looking misses	Hard for a snippet reader. Bounded with whole-ad inputs.
Unverifiable truth. "Clinically proven" is a violation only if substantiation is missing, which text cannot show.	Not measurable	Hard
New-claim drift. The sealed set shares no 6-word phrase with training. Mini alone scores 92.1% on 2021 to 2026 FDA letters there, against 94 to 97% on the earlier random hold-out.	Roughly 2 to 5 points in-distribution; out of distribution unmeasured	Bounded. Needs refresh.
Adversarial rewording. Minimal pairs (negated, disclaimed and neutral rewrites) are in training. No adversarial attack set has been run.	Unmeasured	Hard in principle
False-positive budget. Recall is a function of tolerated false positives (section 2). Going from 5% to 10% buys 10 raw points for the mini, then returns shrink fast.	See the curve	Hard tradeoff

At the shipped 7.1%, a clinic with 100 clean sentences sees about 7 flagged. At 15% it sees 15, which is not a product anyone will use.

5. What more content would add

All gain figures are **estimates** from the arithmetic in sections 2 and 3, not forecasts. Real violation claims on file: about 5,865 FDA 2021 to 2026 claims, 1,737 pre-2021, 574 OPDP, 543 ASA, plus small FTC, ASAI and Canada sets, under 10,000 in all. Free, robots-compliant US regulator sources are mostly drained.

Source	Volume	Legal route	Estimated gain	Why
FDA warning letters that failed to download	302 of 1,156 selected letters (26%)	Public domain; retry <code>fetch.ts</code>	+0.2 raw	2021 to 2026 letters are already at 92.8%, agreed 100%.
Complete OPDP untitled letters plus the cited promo pieces	113 letters listed, 106 fetched; sponsor media not downloaded	Letters public domain; sponsor media is copyrighted, needs owner decision	+1.5 to +2.5 raw if used by a whole-ad model	68 OPDP misses are 6.6% of raw. A third recovered. Needs the model change in section 6 too.
FTC case files and complaints	ftc.gov returned 403; only 16 claims via Federal Register	An approved route (documented request, or owner sign-off on the access terms)	+0.3 raw, more for the product	Sealed has no FTC records, so the benchmark will not move. Disease and supplement claim wording would generalise better.
NAD decisions	1,941 violation and 177 compliant summaries were collected, then deleted	Licence from BBB National Programs only	+0.5 to +1 on superiority and substantiation	Our weakest-supported category, but sealed barely tests it.
Platform ad-rejection logs from pilots	Hundreds to thousands per pilot	Customer agreement	Cannot be forecast	The only source of real in-distribution positives and negatives.
Real customer copy with labels	Target 1,000	Customer agreement	Not forecastable. Gives the first real false-positive rate.	First measurement of the thing we cannot measure today.
Human-labelled sample	200 to start, 1,000 to train on	Reviewer time	Measures the ceiling; +1 to +2 raw on the 144 disputed records and noisy targets	Label noise is the largest unknown.

Evidence that more of the same data has diminishing returns. Round 3 grew training from 7,247 to 28,224 examples (about 3.9 times) and added 1,732 agreed real violations. The mini moved from 76.6% to 80.6% raw. That change also included a different recipe (16 epochs, teacher v3), so the data share cannot be separated. The same distillation gained about 6 points from 6 to 12 epochs (75 to 81% on the tune half). Four times more would need about 100,000 examples, and we lack the real violations to fill them. Synthetic violations are Claude-written and Claude-verified, which may inflate in-style performance.

6. What more model work would add

Estimates, with reasoning.

Work	Expected gain at equal FPR	Cost	Reasoning
Larger student (60 to 100 MB)	+1.5 to +2.5 raw, +1 agreed	GPU hours; a bigger download	The teacher-student gap is 3.3 raw (83.9 against 80.6) and 1.8 agreed. A student will not reach the teacher.
Ensembling	+0 to +0.5	Compute at inference	Measured: union is at the same place as lower thresholds on the mini.
Retrieval of rule and citation context	+0.5 to +1 on implied claims	A retrieval layer on the edge	The 9 implied-claim misses may be fixed by rule text. Mostly improves explanations.
New and tighter rules for the 21 plausible misses	+0.5 to +1 raw	Hours of work	"Approved by the FDA", "preferred option", dosing for an unnamed symptom. Rules were built from these sources, so a rule gain on sealed would be tuned on it: do not count it.
Whole-document model (sees the full ad)	+1.5 to +2.5 raw on OPDP	Large: needs ad-level training data we do not have	Fair balance and omission are visible only across the full text.
Active learning from pilot overrides	Unforecast, likely the best long-run lever	Product plumbing and customer consent	Real distribution, real labels from reviewers.

7. Ranked plan

#	Action	Effort	Cost	Expected gain (estimate)
1	Human audit of 200 sampled labels, including 100 of the 322 non-claims	1 day	Reviewer time	Sets the true denominator; may move raw from 84.4% toward 90% by relabelling non-claims. Changes the measurement, not the product.
2	Real customer copy from 30 to 50 pieces; measure false positives	1 to 2 weeks, depends on customers	None	Replaces the 800 AI-written texts as the false-positive reference.
3	Larger student distilled from claims-v3	2 to 3 days	GPU hours	+1.5 to +2.5 raw, +1 agreed
4	Rule patches for the 21 plausible misses, scored on a new unseen set, never on sealed	1 day	None	+0.5 to +1 on a fresh set
5	Whole-ad OPDP model with sponsor materials	3 to 6 weeks	Needs an owner decision on copyrighted media	+1.5 to +2.5 raw on OPDP
6	FTC case files by an approved route	1 to 2 weeks	Counsel and owner OK	+0.3 raw, better US generalisation
7	NAD licence	Negotiation	Licence fee, unknown	+0.5 to +1 on substantiation claims
8	Active learning from pilot overrides	After pilots	Product work	Unforecast

8. The ceiling, stated plainly

- **Agreed-only: about 99% at 5 to 10% false positives.** Today 98.2% (mini plus rules), 98.6% (teacher plus rules), 99.1% (union). The 5 to 10 records still missed amount to about 1 point. One record is 0.18 points and the 95% interval is about 1 point either way, so work here is not measurable on this set.
- **Raw (estimate): about 88% at 7%, 92 to 94% at 10%, 96 to 97% at 15 to 20%.** Today: 84.4% at 7.1%, 91.9% at about 10%, 95.5% at 15%. No credible path to 100%. 141 of the 162 raw misses are text that neither pass called a violation, and the only way to flag more of it is to accept more false positives on real copy. A larger student is worth about 2 points at equal false positives. Past that, raw points are bought with false positives. A human relabel of the non-claims would raise the reported raw figure by changing the denominator, not the product.
- **Hard limits:** disagreement between regulators and labelers, snippets without the ad around them, claims whose truth we cannot check, and the false-positive tradeoff. **Just work:** a larger student, the 21 plausible misses, the failed letters, and a whole-ad OPDP model.
- **Not yet known, and the biggest risk to the product:** the false-positive rate on real small-business copy. No score in this report answers it.

Sources: `docs/model.md` (round three); `rules/corpus/model/r3/*.log` and `r3/cache/label-{es,eh}.jsonl`; `scripts/model/eval-fresh.py` (the matched-false-positive and per-bucket figures come from a throwaway script that reuses its functions, not committed); `scripts/harvest/eval-real.ts --limit 600`; `rules/corpus/harvest/*/MANIFEST.md`. The agreed denominator comes from AI labels: a bound, not proof.