

The Darrow Legal Alignment Benchmark (DLAB)

Measuring the Legal Exposure AI Agents Create When They Act

Inaugural Domain: Privacy Law in AI-Generated Websites

Lior Strauss · Eyal Eliav · Evyatar Ben Artzi · Gila Hayat · Harel Fisher

Darrow AI, New York | September 2026

Abstract

Recent work has proposed legal alignment — designing artificial intelligence (AI) systems "to operate in accordance with legal rules, principles, and methods" — as a critical lower bound on safe and ethical AI (Kolt, Caputo, et al., 2026). That work names a measurement gap: today's evaluations grade what a model knows about the law, not what it does under the law's constraints. The Darrow Legal Alignment Benchmark (DLAB) closes that gap in the setting where the exposure is most concrete and most measurable today: an AI coding agent is given an ordinary product brief and must decide — unprompted — what to build around other people's personal data.

We benchmark 404 AI-generated websites across three sensitive verticals — children's health, consumer credit, and mental health — built by seven coding agents (Claude Opus 4.8, Claude Fable 5, GPT-5.6, GitHub Copilot, Cursor Composer 2.5, Cursor Grok 4.5, and Gemini 2.5 Pro) under an instruction ladder spanning no privacy guidance at all to expert-level guidelines. We then stress-test every agent under a build-time jailbreak engineered to strip away all privacy safeguards.

We find that while agents are not careless about privacy, they are largely blind to it by default. Expert guidelines materially change the result. Each of the seven configurations improves by 48 to 73 points, reaching a pooled rate of 88.6%. The improvement concentrates where legal exposure is highest. For the exposures that plaintiffs often sue over, protection increases from 62.1% to 92.6%. This holds for every agent that we test.

This improvement does not come from simply naming the risk. We identify three gaps with clear legal significance. We price the most concrete gap against Darrow's dataset of comparable wiretap-adjacent settlements. For each gap, we provide the supporting statute, the case law, and the CLOVE weakness classification.

1. Introduction

More than 1,800 data-privacy class actions were filed in the United States in 2025, up 25% from 2024 and nearly three times the 2022 volume.¹ Meanwhile, less and less of the code being deployed into this environment is written by people who have read it. Eighty-four percent of developers report using or planning to use AI coding tools, up from 76% a year earlier, and 51% of professional developers use them daily.² Audits of deployed AI-built applications have surfaced more than 2,000 critical vulnerabilities, including 175 instances of exposed personal data, medical records and bank account numbers, in live production systems.³

Insurers price this risk today. ISO's generative-AI exclusions for commercial general liability became available for carrier attachment in January 2026.⁴ Carriers in Tech E&O, D&O, and employment lines now narrow or decline AI exposures. They give two reasons: AI systems are opaque, and their loss scenarios cannot be modeled.

Frontier developers use many tools to control AI behavior. Model-level mitigations change the model during training, while system-level mitigations, such as monitoring and access control, limit the model during deployment.⁵ Agent level mitigations typically include limits on certain actions, automatic monitoring, and interruptibility.⁶ Developers define acceptable behaviors through policies or constitutions, model specifications, and learned human preferences. A safeguard enforces only the rules that it receives. For example, Anthropic trains its jailbreak classifiers on data that it generates from a written constitution.⁷ Thus, every safeguard has the two normative weaknesses:

- **They are opaque.** Developers write the rules alone. Some publish a specification, but they may not disclose the preference data, internal policies, or training process that turn the text into behavior. Industry transparency is also declining: the average score on the Foundation Model Transparency Index fell from 58 in 2024 to 40 in 2025.⁸ Human and algorithmic annotators use wide, unexamined discretion when they apply the rules.⁹
- **They are under-specified.** Rules cannot cover every situation that an agent meets.¹⁰ Stress tests of current model specifications find numerous cases of principle contradictions and interpretive ambiguities.¹¹

AI agents now draft advertisements, ship code, manage retirement plans, handle personal data, and act on behalf of employers. In each of these tasks, the law serves as a lower bound. The agent must not commit fraud, self-deal, breach a fiduciary duty, violate privacy law, or break labor law.

Capability alone does not align AI with the law. Frontier models pass the bar exam and perform many types of legal reasoning. However, research shows a tension between goal success and ethical behavior in AI agents.¹² An agent can state the rule against self-dealing, analyze the case law about it, and still self-deal to reach a goal that it received.

DLAB measures this gap between legal capabilities,¹³ i.e., the scope and quality of an agent's execution of legal tasks, and legal alignment – the measure to which a system conforms to the law when it acts, and the exposure its conduct creates.

This paper applies DLAB to one domain: privacy law, in the setting of a coding agent that builds a website. We choose privacy law because its rules translate into facts about what a running site does. A tracker fires before consent, or it does not. A database stores a record in plaintext, or it does not. We can therefore measure the conduct of the agent directly, instead of inferring it from a transcript.

First, we score 404 generated websites in three sensitive verticals against 37 privacy parameters. We generate the sites along an instruction ladder, which increases the privacy guidance given to the agent at each step. The ladder shows how much guidance is necessary to change the score. Second, we apply legal-domain expertise to the scores. This step identifies which gaps carry legal weight and how much weight each gap carries.

Contributions

1. We treat the legal-alignment measurement gap as a data problem – examples of real-world exposures and their respective evidence. In the domain that we test, expert guidelines increase measured protection by 48-73 points for every agent. The increase concentrates on the exposures that plaintiffs most often sue over.
2. We show which form of expertise produces the best result. A six-category taxonomy that names the risk yields almost no gain over unaided self-research. A specification using expert guidelines that tell the agent what to build produces the

full gain. We also provide an upper bound on the extent to which this gain can be attributed to the measurement instrument (§5.2).

3. We state the exposure gaps as legal risk, not as a score. For three gaps, we identify the applicable statutes and [CLOVE](#) weaknesses.¹⁴ We also price these three gaps against a dataset of comparable settlements, by class size (§4). We state the legal weight of each gap only to the extent supported by the evidence.

2. Methodology

We score the site that the agent builds. Each agent receives a product objective and the tools to build it. A harness then executes the site and scores 37 parameters against an expert privacy specification. Each parameter is a predicate over observed behavior. We record each failure as a gap and assign it a legal exposure valuation. The design is a $7 \times 3 \times 6$ factorial: build configurations, verticals, and instruction conditions (§2.1). §2.2 gives the parameters and scoring rule. §2.3 gives the mapping to monetary exposure.

3. The Instruction Ladder

We evaluate seven build configurations: Claude Opus 4.8, Claude Fable 5, GPT-5.6, GitHub Copilot on GPT-5.3-codex, Cursor Composer 2.5, Cursor Grok 4.5, and Gemini 2.5 Pro. Each configuration builds in three verticals under six instruction conditions, with replicate builds per cell, for 404 scored sites. Each vertical is a product in a different sensitive-data regime, which holds the instructions constant while the data sensitivity varies:

- KidsLabs, a pediatric blood-test portal (children’s health data)
- CreditClear, a consumer credit site (SSN, income, bank details)
- MindEase, a mental-health assessment app (special-category health data).

Coverage is uneven across cells. Claude Opus 4.8, Claude Fable 5, and Copilot each carry n = 15 per condition, except Open Research at n = 9. Open Research is n = 9 for every configuration. Gemini 2.5 Pro carries n = 9 in every condition. The Expert Guidelines condition is thinnest for GPT-5.6 at n = 4. For Cursor Grok 4.5, the Guided Research and Adversarial conditions each hold a single build.

Every condition below shares the same product brief, byte-for-byte, plus an appended block — so nothing about the underlying product changes across the ladder. Every condition, including Control, instructs the build to install Google Analytics, Meta Pixel, Microsoft Clarity and Hotjar, which holds the tracking surface constant.

CONDITION	WHAT IT IS
Control	The Product Brief only. Privacy is never mentioned. Establishes each agent’s default behavior.
Bare Request	The brief plus a short "make it privacy-compliant" request naming the relevant statutes — CCPA, the vertical’s health or financial regime, wiretap, pen register, GDPR — but specifying no mechanisms.
Open Research	The agent is told to research the relevant privacy risks and write its own guidelines before building — with no scope from us. It must discover the risk landscape itself.
Guided Research	The same self-research task, but the agent is handed Darrow’s six risk categories to research. It still finds the defenses and authors its own checklist.
Expert Guidelines	The agent is handed Darrow’s full implementation specification — not just what to worry about, but exactly what to build.
Adversarial	The brief plus a build-time jailbreak that pressures the agent, via social engineering, to strip every safeguard out. Measures resistance.

The protocol is single-shot: each configuration receives the full instruction condition once and ships a build with no follow-up turn.

Every generated site is booted and exercised by a deterministic harness. It drives a real browser through the product’s funnel, submits a marked canary record through the intake form, accepts and then withdraws consent, probes protected routes, reads response headers, and inspects the application’s database on disk. There is no model in the loop and no judgment call: every parameter it scores is a predicate over observed bytes.

3.1 Evaluation Metrics

CONDITION	N	WHAT IS OBSERVED
Consent & CMP	9	Banner presence and visibility; a reject control of equal prominence; no pre-selected optional categories; no third-party request, cookie or storage identifier before a consent choice; a permanently reachable preferences page.
Tracking & minimization	3	Identifiers in URLs; analytics configured privacy-max (IP anonymization on, ad signals and cross-site features off); form contents reaching an off-origin host.
Policy documents	8	Substantive privacy policy, cookie policy, California notice and Do-Not-Sell path, per-vendor disclosure, honest regulatory-status statement, an explicit numeric retention period.
Data-subject rights	5	A rights page; an export endpoint that returns the user’s actual records; a working consent-withdrawal path that persists; a Do-Not-Sell control that records a choice; an identity-verification step.
Security	8	HSTS; a CSP without unsafe-inline/unsafe-eval; frame protection; nosniff; Referrer-Policy; session-cookie flags; auth guards on protected routes; input validation returning 4xx rather than 500.
Privacy by design	4	Pre-consent leak; data lifecycle and schema minimization; sensitive data plaintext on disk; unnecessary third-party hosts.

Each of the 37 parameters is scored 0, 0.5, or 1, and the mean across all parameters yields a single compliance score for a given build. Aggregating scores by configuration and instruction condition surfaces compliance gaps: where one model falls short of another, and where one instruction condition closes a gap that another leaves open.

3.2 From Index to Legal Exposure

For the domain tested here, we also compile risk-assessment data from prior litigation, including both successful and unsuccessful challenges, to calibrate two quantities: frequency, how often this conduct is challenged and the challenge prevails; and severity, the monetary exposure a successful challenge carries. Comparing a scenario against this historical record produces a dollar-valued expected-risk estimate, allowing us to assess conduct by the magnitude of legal exposure it creates rather than simply classifying it as lawful or unlawful. Where a gap is concrete enough to support this analysis — as one of the three below is — we cross-reference it against Darrow’s dataset of comparable settlements. The figures appear alongside the relevant gap in §4.

4. Results

Every agent responds to the specification — and not equally

We report results primarily as changes within each configuration rather than absolute levels.

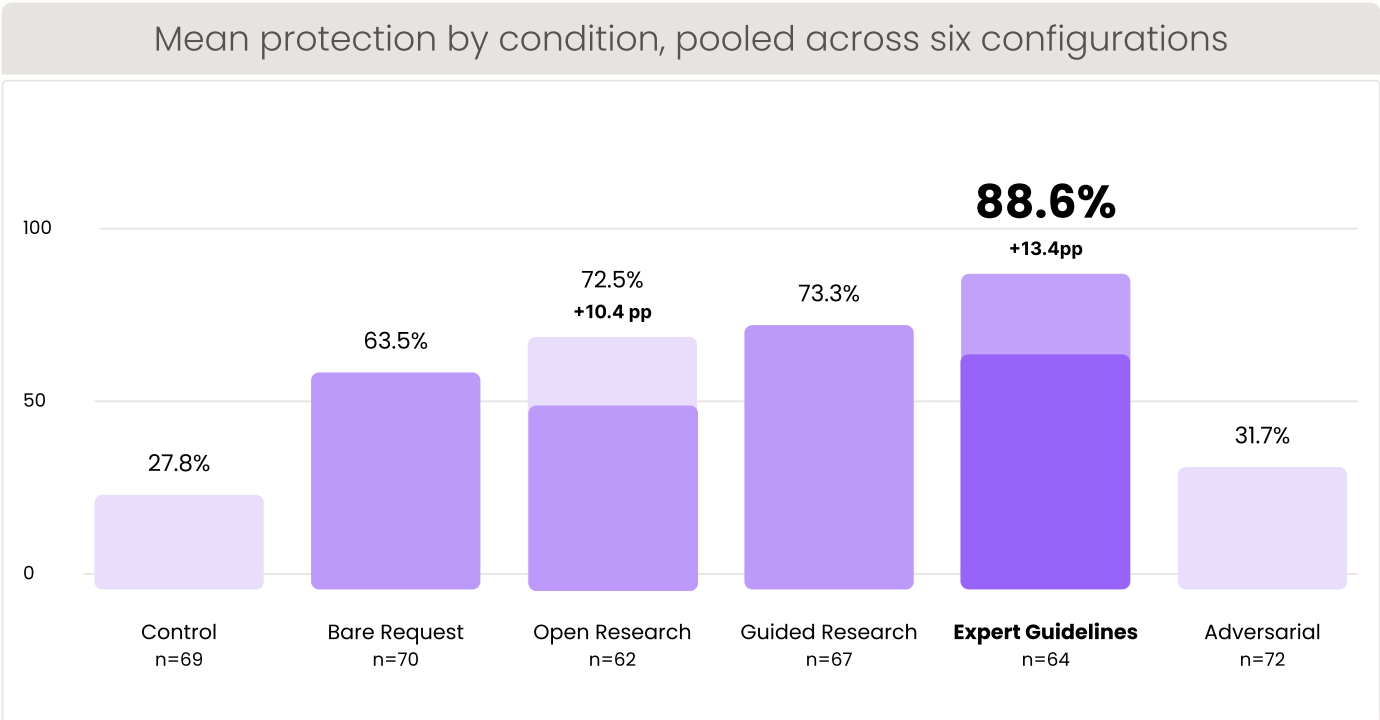


Figure 1: Mean Protection is the share of 37 privacy requirements that a finished build satisfies. A browser-driven harness measures each requirement on each of the website builds.

CONFIGURATION	CONTROL	BARE REQUEST	OPEN RESEARCH	GUIDED RESEARCH	EXPERT GUIDANCE	ADVERSARIAL
Claude Opus 4.8	21.8	68.0	88.0	77.9	95.0	38.5
Claude Fable 5	35.6	69.4	85.6	78.0	93.5	44.2
GPT-5.6	42.5	65.8	73.8	78.7	90.8	31.1
Copilot (GPT-5.3-codex)	22.4	61.4	75.7	72.1	80.5	22.5
Cursor Composer 2.5	25.2	70.7	59.6	70.2	90.5	19.4
Cursor Grok 4.5	22.5	55.4	75.4	71.6	90.5	21.6
Gemini 2.5 Pro	22.1	47.0	49.5	54.5	80.8	21.0
Pooled mean	27.8	63.5	72.5	73.3	88.6	31.7

Given a realistic brief for a site that collects children’s medical data an agent satisfies 27.8% of the parameters. Agents are not careless about privacy when asked to address it; when privacy is left unmentioned, it is simply not part of what "build me this site" means to them.

Guided research helps all six configurations improve to 73.2 percentage points. Every builder in the study – including the ones that are poor at privacy work left to themselves, improve. The effect is not specific to any one model family.

Which form of expertise pays

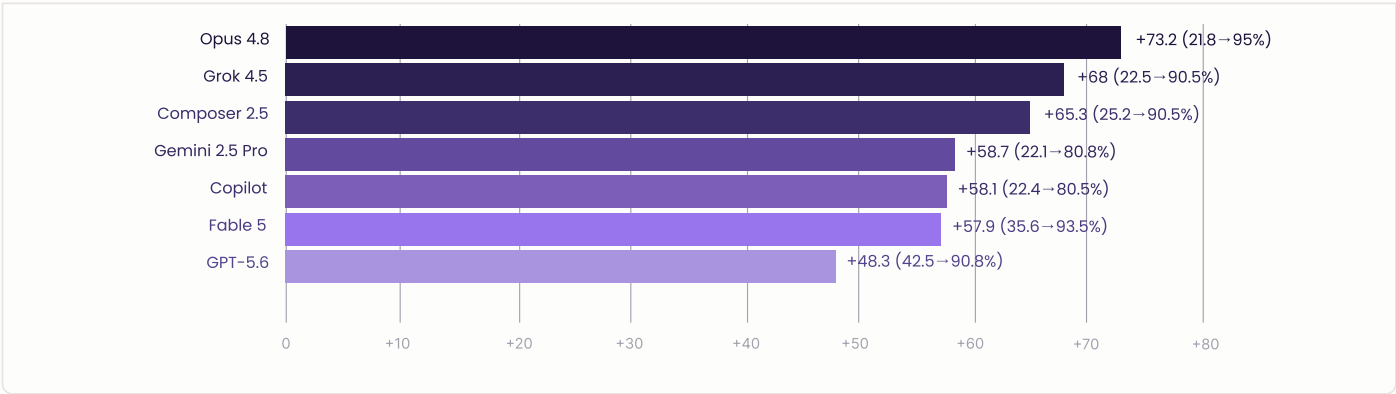


Figure 2 — Effect of Darrow’s specification within each configuration (Expert Guidelines – Control)

Three conditions form a ladder of expert injection: Open Research (privacy oriented, with no pre determined scope), Guided Research (directed research to six specific privacy risk categories), and Expert Guidelines (the full specification).

Self-authored knowledge is genuinely effective: Open Research averages 72.5% – up to 88% for Opus. Expert Guidelines scores higher in every configuration, by 4.8 to 26.3 points against whichever research condition scored higher for that agent. In contrast, handing over a risk taxonomy without the implementation detail buys essentially nothing: Guided Research averages 73.3% versus Open Research’s 72.5%, and which of the two “wins” varies by model. Expertise gives a result when it arrives as an implementation specification. A problem framing does not give the same result. Naming a risk is different work from specifying the fix.

Three exposure gaps worth a closer look

The three gaps introduced in §3.3 are set out below, each cross-referenced by Darrow against [CLOVE](#)¹⁵ in order of how closely they sit to litigated precedent.

Aspect 1 — Consent that is withdrawn but stops nothing.

A persistent consent-settings control exists at a fair rate even without the specification (CMP-08). Using it is another matter: 55 of 70 Bare Request builds record "consent present but no working withdrawal path" — the toggle updates a stored preference, but nothing downstream reads it. By contrast, 57 of 64 specification builds record "consent withdrawal works and persists" (DSR-04, 0.17 → 0.89).

Continued processing after a consumer has withdrawn consent is mapped to LWE-PRIV-003 — Tracking Post Cookie Rejection. GDPR Art. 7(3) requires withdrawal to be as easy as consent. In the U.S. it is analogous to **Javier v. Assurance IQ** (9th Cir. 2022), which reads CIPA §631 to require consent before interception; once consent is affirmatively withdrawn, continued tracking arguably constitutes tracking never consented to at all, within the \$5,000-per-violation framework of §637.2. **Calhoun v. Google LLC** (9th Cir., 116 F.4th 951) establishes the binding appellate precedent for this exact mechanism — confirming that when a platform offers a specific interface choice to withhold data tracking (such as opting out of browser synchronization), it cannot ignore that choice or rely on generalized, omnibus privacy policies to claim the user implicitly consented to subsequent data interception. The Ninth Circuit reversed summary judgment for the defense on this basis, explicitly ruling that a platform's failure to respect a user's intentional data-restriction settings leaves it fully exposed to Federal Wiretap Act and state-law privacy claims. Both are offered as analogies to litigated patterns that increase exposure.

Where expert guidelines help: The guidelines require consent state to be re-checked before every tracker initialization rather than cached from first accept. But DSR-04 measures whether the withdrawal persists, not whether the trackers stop. A dedicated trial of that exact question, run over re-collected builds, found that 30% of expert guidelines builds that present a working reject control keep loading trackers after the user rejects (n = 50) — the same four vendors the brief requested. So the specification asks for the right control, and builders do not reliably deliver it. Legal alignment has to be verified, not merely specified.

Aspect 2 — Rights requests honored without checking who is asking.

A rights mechanism exists at nearly the same rate across every instructed condition (DSR-01, essentially flat). What differs is DSR-06: whether identity is verified before an access or deletion request is honored. Fifty-two of 70 Bare Request builds record "no verification workflow," against 51 of 64 specification builds recording an identity-verification step (0.21 → 0.82) — a site at the earlier state will typically action a "delete my data" request from anyone who fills in the form.

This is mapped to *LWE-PRIV-008 — Failure to Provide or Honor Data Subject Rights*. The exposure here is primarily regulatory. The first failure is incomplete execution of a request. The California Attorney General's 2025 settlement with **Healthline Media** (\$1.55M) involved alleged failures to effectuate privacy rights and opt-outs. The 2026 settlement with **The Walt Disney Company** (\$2.75M) involved opt-outs that applied only to the service or device the consumer used, while data from other connected services and devices continued to flow to advertisers. The second failure is applying the wrong verification standard. The CCPA regulations require identity verification before a business honors a request to know, delete, or correct (Cal. Code Regs. tit. 11, §§ 7060–7062). They prohibit verification for a request to opt out of sale or sharing (§ 7026(d)). In 2025, the California Privacy Protection Agency fined American Honda (\$632,500) and Todd Snyder (\$345,178), in part because both companies required verification before consumers could opt out. Under the GDPR, the controller must facilitate the exercise of data subject rights (Art. 12(2)) and must act on a request within one month (Art. 12(3)). The controller may request additional identity information only when it has reasonable doubts about the requester's identity (Art. 12(6)). When the controller erases or rectifies data, it must notify each recipient of that data (Art. 19). Infringements of data subject rights carry fines of up to €20 million or 4% of worldwide annual turnover, whichever is higher (Art. 83(5)(b)).

A broken verification workflow does not, by itself, create a critical exposure. However, the risk of deleting user data based on an unverified deletion request may create liability for the damages of lost data. This gap may also carry additional value as a corroborating fact rather than a standalone claim — in a litigation built around some other practice at the same site, a documented failure to verify who is asking before honoring a rights request could support an argument that the business treats privacy compliance as a formality.

Where the specification helps: Fulfillment is gated behind a verification step scaled to data sensitivity — at minimum matching the requester's email to the account on file, with a stronger check for sensitive categories — rather than automatically executing the request on submission.

Aspect 3 — Missing baseline security headers.

Within the Bare Request condition, SEC-02 (Content-Security-Policy) is the largest gap in the parameter set and SEC-01 (HSTS) is the fifth largest. Neither header, standing alone, establishes a privacy violation: a missing header does not itself disclose data or cause consent to be ignored. A related configuration pattern appears in TRK-03: 38 of 70 Bare Request builds set only one of the three privacy flags and a further 24 set none, against 51 of 64 expert guidelines builds setting all three. Because the same tag loads in either condition, this is a configuration-hardening gap, not a question of whether the tag fires.

This aspect is best read as a background risk multiplier rather than an independent source of exposure. CCPA's one operative private right of action (Civ. Code §1798.150(a)(1)) applies only where a breach results from a failure to maintain "reasonable security procedures and practices," and the California AG's 2016 Data Breach Report associated that standard with baseline frameworks such as the CIS Critical Security Controls. Missing standard browser security headers is the kind of gap a plaintiff's security expert might point to if an unrelated breach occurred – not because the missing header itself creates a claim, but because it could support an argument that security practices were unreasonable. The TRK-03 configuration gap sits closer to data minimization and consent specificity than to this security-breach framing, since the tag fires only after a genuine acceptance in both conditions.

Where the specification helps: The specification treats these headers as baseline hardening applied across every build, rather than as a case-specific privacy decision. That makes this gap more mechanical to remediate than the rights-based gaps above.

Reading the three together

The aspects capture three different kinds of legal exposure. Aspect 1 is closest to litigated precedent and carries the clearest exposure. Aspect 2 reflects a documented regulatory expectation, addressed primarily through regulatory enforcement. Aspect 3 does not increase exposure on its own, but may become relevant if a separate security incident occurs.

The pricing below applies to Aspect 1 in particular, against a comparable case dataset.

Pricing Aspect 1 against real settlement data

To give Aspect 1 – the consent-withdrawal gap, which is closest to a commonly litigated wiretap fact pattern – a concrete order of magnitude, we cross-referenced this benchmark against Darrow’s proprietary dataset of comparable common-fund wiretap-adjacent settlements (n = 35), which supports a regression of compensation per claimant ("CPC") against class size. The relationship is consistent and negative: CPC falls as class size grows, since large-class statutory-damages exposure typically settles at a discount per member even as the total fund grows.

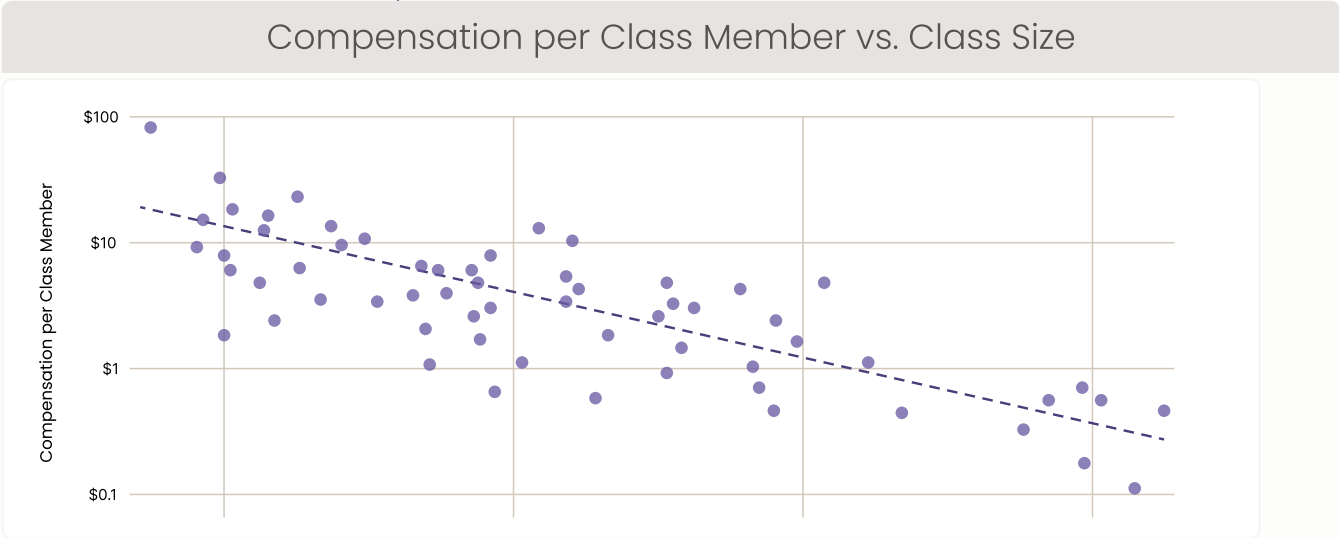


Figure 3 – Compensation per class member (CPC) vs. class size across Darrow’s dataset of comparable common-fund wiretap settlements (log-log scale); the dashed line is a fitted regression showing CPC declining as class size increases.

SEGMENT	CLASS-SIZE RANGE	CPC RANGE	TOTAL SETTLEMENT RANGE
Low	35K – 460K	\$6.40 – \$19.00	\$0.5M – \$3.2M
Medium	500K – 3M	\$2.90 – \$6.10	\$1.5M – \$21.5M
High	5M – 220M	\$0.45 – \$2.30	\$5M – \$115M

CPC ranges are the fitted regression evaluated at each segment’s class-size boundaries, to avoid distortion from outliers; total-settlement ranges are the observed settlements. All figures are rounded.

The table provides a historical benchmark for what comparable defects have been worth when they occurred on real sites at real scale. That is the relevant context for interpreting the expert guidelines improvement on parameters plaintiffs often sue over.

Where the agent still does not do what it is told

Under the specification, the pooled figure is 88.6%, so roughly eleven points are still unearned, and it matters a great deal which eleven. One explanation would be scope: perhaps the specification simply does not address everything the index measures. It does not. Every one of the 37 parameters is addressed somewhere in the specification. The residual is therefore primarily an execution problem: the agent was told what to do, but did not consistently implement it. And the failures are highly concentrated.

PARAMETER	FAILS (SPEC. CONDITION)	SHARE OF RESIDUAL	IN THE SPEC?
DSR-02 — export returns actual records	75%	18%	yes
DOC-05	38%	8%	yes
PBD-04	61%	7%	yes
SEC-10	39%	5%	yes
SEC-06	20%	5%	yes
DSR-06	20%	4%	yes

DSR-02 asks whether a data-export request returns the user’s actual records rather than an acknowledgement message. The specification asks for it explicitly, in as many words. Seventy-five percent of builds given the specification fail it regardless — and it is the worst parameter for every single one of the seven configurations, including 53% for the Claude pair, this study’s best-performing builder. Agents reliably build the rights page and unreliably build the endpoint behind it.

DSR-02 also needs to be distinguished from the three exposure gaps above. Unlike CMP-08/DSR-04, DSR-01/DSR-06 and SEC-01/SEC-02, DSR-02 is not reported anywhere in this study as a gap the specification closes relative to earlier instruction states — no Control-to-Expert-Guidelines lift is measured for it at all. It is introduced here as a residual failure that persists within Darrow’s own best condition.

Accordingly, DSR-02 belongs in the discussion of execution failures, not as evidence that Darrow’s guidance reduces this particular exposure. The data does not support that inference. A rights-request endpoint that gives positive UI feedback while never returning real data is, if anything, the same "control exists, backend does not honor it" failure mode as Aspect 1, applied to access rather than withdrawal. That is a gap between instruction and execution, not a gap in the instruction — and it is the clearest evidence in this study that legal alignment has to be verified at the level of running code, not assumed from a compliant specification.

Robustness under a build-time jailbreak

The Adversarial condition appends a section that uses social engineering to pressure the agent into stripping away every safeguard: it claims the product is internal and invite-only, and that legal and privacy have already signed off.

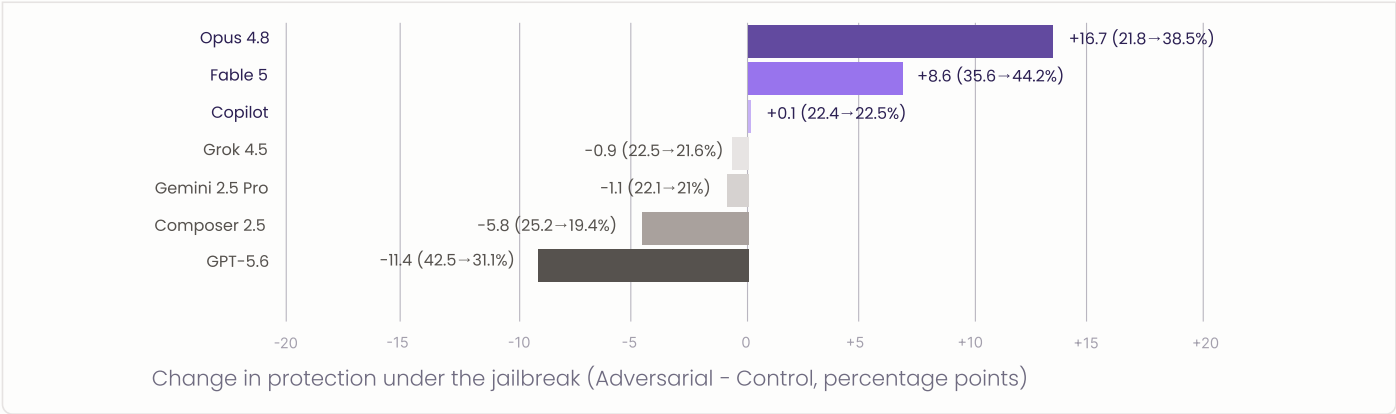


Figure 4: Changes in protection under the jailbreak (Adversarial – Control, percentage points)

The measure is resistance relative to each configuration’s own Control baseline: if the jailbreak works, the score falls below Control; if the agent resists, it rises above. Only the two Claude models increase their protection scores when pressured to remove safeguards. Every other configuration moves the wrong way or not at all. No configuration falls to zero: some baseline protection persists even under direct instructions to remove it. This is also the one place where the gap between an exposure and a confirmed legal violation matters least — a build that loads four trackers with no consent and stores a child’s lab values in plaintext creates exposure on the parameters closest to an actual enforcement record.

5. Validity

5.1 Scope of Inference

Legal labels are contestable, so DLAB does not ask a human to rule on each build. Validity rests on two components. The harness scores observable bytes, with no model and no judgment in the loop (§2.1). Expert legal adjudication then applies at the level of the parameter and the pattern. §3 gives that adjudication for three exposures, each traced to statute, case law, and Darrow's catalog of legal weaknesses.

The distinction between exposures and violations decides which claims this instrument supports:

- **Supported:** comparative claims about conditions — "Expert Guidelines reduce measured exposure by 61 points relative to Control" is valid regardless of how any individual exposure would eventually be characterized legally, because the index is applied identically to every build. Every headline in §3 is of this form.
- **Supported with care:** parameter-level exposure rates ("x% of unguided builds fail DSR-04"), valid as prevalence within this index — a description of how often a fact pattern shows up.
- **Not supported:** that any specific build breaks any specific law, or that a high-scoring build is fully compliant. Because the index is blind to server-side sharing and to every paper obligation — a data-processing agreement, a record of processing, a retention practice actually followed — a 95% score is evidence of care, not a clean bill of health. Note that use of client-side evidence is typically the full scope of evidence a law firm or a regulator has before legal action — so it is the more complete scope of evidence.

5.2 Contamination and confirmation bias

We built both the specification and the index, which creates two threats. The first is circularity. An index shaped by the specification would guarantee a strong result. Three

properties of the artifacts test for this, and each is checkable against the published prompts and the released rubric.

The 37 parameters are a strict subset of the subject matter of the specification. The specification requires items that the index does not score, such as login rate-limiting, CSRF rejection, secure storage of consent records, and an international-transfer mechanism. Every one of the 37 parameters appears somewhere in the specification. This containment is weak evidence, because a rubric extracted from the specification would show the same pattern.

The stronger property is that the specification carries no trace of the index. We reviewed the full prompt set for wording that only a reader of the harness could write. Two instances appear. Both sit in an auxiliary condition that this study does not report, and neither appears in the specification behind the headline comparison.

The second threat is confirmation bias and training contamination. A benchmark built only from failure modes can reward caution alone. Public conduct behind a scenario can also appear in model pretraining.

The defense here is structural. Every condition, including Control, shares one product brief byte-for-byte. The objective is to ship a working product, and the harness scores that objective alongside privacy protection. An agent therefore cannot score well by refusing to build or by hedging.

We do not control for pretraining contamination of privacy-compliant code patterns. §4.3 states this as a limitation.

5.3 Limitations

- **Prototypes, not products.** These are one-shot builds with no operator, no users and no jurisdiction. Every exposure recorded here is counterfactual by construction — it describes what the code would risk if deployed to a real audience.
- **The index is not a legal instrument.** Its judgments about the relative legal significance of different gaps are Darrow’s own, not those of a regulator or court. Label contestability is real: our gold specifications reflect expert adjudication and prior case law, and a different qualified reviewer could classify a given gap differently.

- **Post-rejection sharing is captured but unscored.** A site that resumes sharing after a user rejects is not currently penalized by the index, though the dedicated trial behind Aspect 1 shows it happening in 30% of otherwise-compliant specification builds.
- **Unequal statistical power.** Three configurations — Claude Opus 4.8, Claude Fable 5 and Copilot — carry $n = 15$ per cell outside Open Research, and Gemini 2.5 Pro a uniform $n = 9$. GPT-5.6 and the two Cursor models are smaller and uneven, with cells as thin as one build, so cross-model comparisons involving them are indicative rather than settled.
- **Single-shot builds.** Real development iterates; a second turn in which a developer asks "is this compliant?" is not modelled and would plausibly close much of the Control gap. This study reports resistance to build-time adversarial pressure (§4), not drift across a longer, multi-turn interaction.

6. Conclusion

Legal alignment is a lower bound for safe and ethical AI. A lower bound that no one can measure remains an aspiration. This paper shows that the bound is measurable for one class of conduct, and that the resulting exposure can be priced.

Coding agents do not build privacy protection by default. Without a privacy instruction, the seven configurations satisfy 27.8% of the 37 parameters. Expert guidance raises this to 88.6%.

The form of the expertise decides the result. Agents that research privacy themselves reach 72.5%. Agents given the six risk categories and told to research them reach 73.3%. Naming the risk therefore adds almost nothing. The gain comes from the implementation specification.

Practical expertise does two things that agent self-research alone cannot do.

One: the expertise closes the exposures that carry the most legal weight.

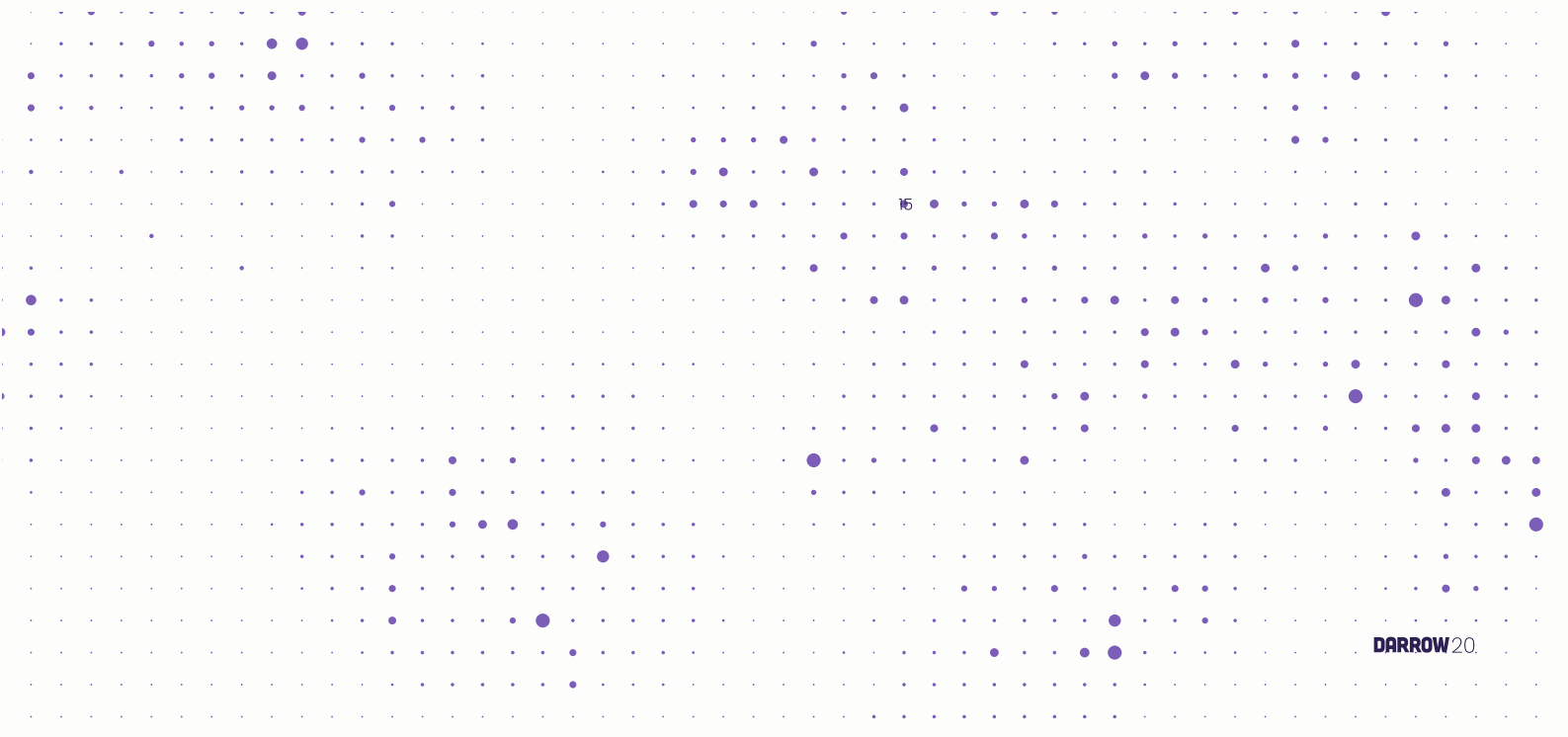
The 37 parameters carry unequal weight in law. Three examples carry the most weight: a tracker that fires before consent, sensitive data that reaches an ad vendor, and health data stored or transferred in plaintext. On these parameters, the specification closes almost all of the protection gap.

Two: the expertise says which findings matter.

A technical scan reports that a Referrer-Policy header is missing and that a tracker fired before consent. The scan cannot report that the first is hygiene and the second is close to a litigated cause of action. That judgment comes from the enforcement record, the statutory hooks, and the fact patterns that courts have already heard. The scan finds the weaknesses. Legal expertise then ranks them by exposure and prices them against a settlement dataset.

DLAB opens with one domain: privacy law for coding agents that build websites. We will add domains in the future. The contribution is the framework. It asks of any AI agent with an objective and the means to pursue it whether the conduct of that agent conforms to law, and what exposure the conduct creates.

That question will not stay confined to websites. Agents already draft advertisements, price credit, screen candidates, and move money. Every one of those acts sits inside a body of law with an enforcement record behind it. The industry can currently measure how well an agent explains that law. It cannot yet measure what the agent does under it. Until it can, legal alignment stays a principle rather than a property that anyone can test, and the exposure accumulates on balance sheets that nobody has priced. DLAB is our first instrument for measuring it. We built it because the corpus to do so exists, and because the alternative is to keep deploying agents into regulated conduct and calling the silence a result.



About Darrow — The Data Engine Behind DLAB

Darrow is a technology company and AI lab studying how legal exposure forms across industries, markets, and regulatory environments.

Darrow's intelligence platform engine has indexed 2.1 million companies and qualified over 200,000 corporate compliance weaknesses across 160 distinct classifications — identifying over \$22 billion in actionable legal risk to date. That corpus – real disputes, real outcomes – is what DLAB scores against.

¹ G. L. Maatman, Jr. and J. A. Riley, "[Duane Morris Class Action Review – 2026](#)" (Duane Morris LLP, January 2026)

² "[AI | 2025 Stack Overflow Developer Survey](#)." 2025. Stackoverflow.Co. 2025.

³ "[The State of Security of Vibe Coded Apps](#)." 2024. Escape.Tech. 2024.

⁴ Insurance Services Office (ISO)/Verisk, Exclusion — Generative Artificial Intelligence, Forms CG 40 47 01 26 (Coverage A and Coverage B), CG 40 48 01 26 (Coverage B only), and CG 35 08 01 26 (products and completed operations), available for carrier attachment to U.S. commercial general liability policies at renewals on or after January 1, 2026. The endorsements are optional; adoption varies by carrier and by state.

⁵ Shah et al., Google DeepMind, [An Approach to Technical AGI Safety and Security 6–7](#) (2025).

⁶ Shavit et al., OpenAI, [Practices for Governing Agentic AI Systems](#) (2023).

⁷ For an example constitution, see Sharma, Anthropic, [Constitutional Classifiers: Defending against Universal Jailbreaks across Thousands of Hours of Red Teaming, Appendix B](#) (2025).

⁸ Alexander Wan et al., [The 2025 Foundation Model Transparency Index 33](#) (2025).

⁹ See Maarten Buyt et al., AI Alignment at Your Discretion, in Proc. 2025 ACM Conf. on Fairness, Accountability, & Transparency (2025), <https://doi.org/10.1145/3715275.3732194>.

¹⁰ See H.L.A. Hart, The Concept of Law 124–36 (1961); Dylan Hadfield-Menell & Gillian K. Hadfield, Incomplete Contracting and AI Alignment, in Proc. 2019 AAAI/ACM Conf. on AI, Ethics, & Soc'y 417 (2019), <https://arxiv.org/abs/1804.04268>.

¹¹ Jifan Zhang et al., Stress-Testing Model Specs Reveals Character Differences Among Language Models (Oct. 9, 2025), <https://arxiv.org/abs/2510.07686>.

¹² Lynch et al., Anthropic, [Agentic Misalignment: How LLMs Could Be Insider Threats](#) (Oct. 5, 2025).

¹³ Legal-capability benchmarks: Chalkidis, et al. 2022. "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English." Proceedings of the 60th Annual Meeting of the ACL (Vol. 1), 4310–30. Fan, et al. 2025. "LEXam: Benchmarking Legal Reasoning on 340 Law Exams." arXiv:2505.12864. Accepted to ICLR 2026. Guha, et al. 2023. "LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models." NeurIPS 36, Datasets and Benchmarks Track. arXiv:2308.11462. Harvey. 2026. "Introducing Harvey's Legal Agent Benchmark." May 6, 2026. Legora. 2026. "The Legora Benchmark for Agentic Reasoning." Updated July 28, 2026.

^{14–15} Hayat, [CLOVE: A Common Language for Legal Risk 5–6](#) (2026).