Detection Objection: Documenting, Calibrating, and Detecting the Rise of AI-Assisted Text in U.S. State Courts

JAVOKHIR ARIFOV, ALEXANDER SPANGHER, JONATHAN HENNESSY, DIYA KARAN, EMAAN HARIRI, DANIEL E. HO,
Stanford University, USA

As reports of legal hallucinations flood the headlines, there is growing concern that litigants increasingly use generative AI to draft court filings. Yet, there is limited robust evidence on AI-assisted text in American courts. State court records are difficult to collect, and AI detectors are poorly validated on legal text. We offer the first study of generative AI prevalence across four state trial court systems. First, we show why legal text is hard for detectors: roughly a quarter of tokens in legal filings are boilerplate, which dilutes the detection signal. Detectors are also highly sensitive to modeling settings. Across reasonable settings for one detector, the share of filings flagged as AI-assisted ranged from under 5% to almost 50%. Second, we introduce PhantomDocket, a dataset of 980 filings *known* to have used generative AI, traced from rulings in Charlotin's AI Hallucination Cases Database. Third, we use PhantomDocket to benchmark detectors and select their settings. Fourth, we introduce ScoreMix, a prevalence estimator that leverages PhantomDocket and is less sensitive to detector settings: under two candidate settings, its 2026 estimates differ by 0.6–5.6 percentage points. Last, we apply ScoreMix to 127,518 complaints from San Francisco, Santa Clara, Oklahoma, and Tulsa counties. Estimated prevalence is near zero before 2023 and rises sharply thereafter, varying across courts, case types, and representation status. In 2026, we estimate that 36% of complaints in San Francisco and 38% in Santa Clara show AI-assisted drafting, compared with 16–17% in the Oklahoma counties. Among pro se complaints in Santa Clara, the estimate reaches 72%. Because PhantomDocket captures AI use severe enough to produce hallucinations, these estimates may be conservative. Our findings inform both the rise of AI in law and the methods used to detect it.

Additional Key Words and Phrases: AI detection, large language models, court filings, legal text, prevalence estimation, calibration

1 Introduction

The rise of AI usage in writing tasks has been documented across a range of industries [26, 27], including law [2, 49], where AI usage has resulted in notable errors and lapses in human judgement

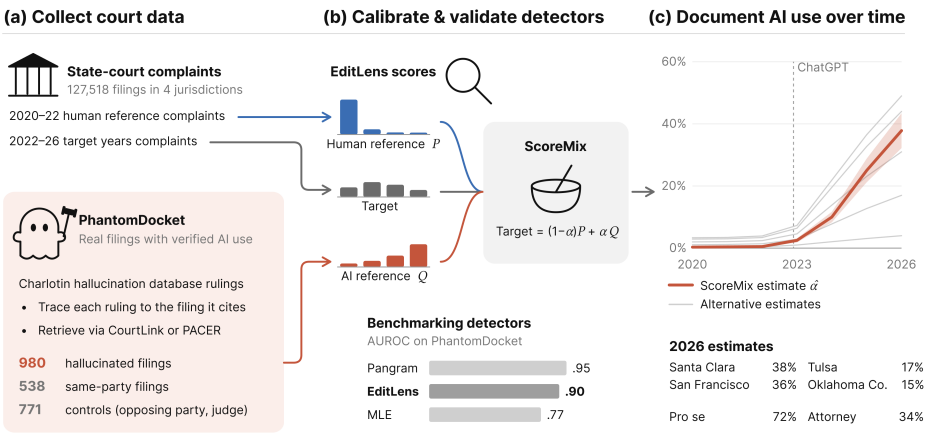


Fig. 1. **Overview of the AI prevalence data, validation, and estimation contributions.** We collect state trial court complaints, construct PhantomDocket from filings with verified AI use, benchmark and calibrate AI-text detectors, and apply ScoreMix to estimate AI prevalence over time.

(e.g. [52]). Many of these are documented in Charlotin's AI Hallucination Cases Database [7]. This has led to the rise of many AI products tailored to legal research and writing [29] and also to considerable internal deliberation within the legal community around the appropriateness, role and extent of AI in education, research and writing [10]. Researchers have identified impacts in court rooms [35] and emerging work has claimed that AI usage is rising to encompass close to 18% of cases in some districts, and is implied to explain a rise in self-represented (pro se) litigants [8, 45]. If true, the impacts of this could be considerable: AI assistance could lower barriers to legal expertise and access to the justice system, but it could also strain a court system that is already at capacity by increasing the volume of cases, the length of documents, and likelihood that citations and facts are fictitious.

Measuring this rise is difficult. Survey evidence is limited and exhibits high variability [2, 46, 49]. Increasingly, there has been interest in relying on AI detection methods – which purport to measure the likelihood that a document was generated with AI – but available approaches do not reliably generalize across domains, language models, and generation settings [13, 19, 51]. Detectors can falsely flag human writing [25], ignore AI text after rephrasing [23, 42], and miss AI-assisted text [48]. Even then, their outputs depend on choices about which textual units to score, how those scores are aggregated, and where detection thresholds are set. We show that these choices are consequential: across 216 settings of one detector, the share of filings across our four counties detected as using AI runs from under 1% to over 99%. We also argue that characteristics of legal text (boilerplate, jargon, and unknown usage patterns) make AI-detection particularly challenging in this domain.

No prior work systematically tests different modeling settings against verified validation datasets in the legal domain. To quantify how AI detectors perform in legal practice, we introduce a novel dataset called PhantomDocket, which consists of 980 full documents with known hallucinated citations derived from the Charlotin database, 538 full documents by the same party but unconfirmed to have used AI, and 771 control documents from presiding judges and opposition parties in the same case. Using hallucinated cases in PhantomDocket as known instances of AI usage, we benchmark multiple AI detection tools, including a distributional maximum likelihood estimation approach (word-occurrence MLE) by Liang et al. [24, 26], EditLens [48], and Pangram [14].

Our primary contributions are as follows:

- We demonstrate theoretical and practical reasons why AI-detection must be approached cautiously in law, and we demonstrate how estimates are sensitive to hyperparameter and reporting choices.
- We develop and publicly release PhantomDocket,[1] which contains legal documents with known instances of AI usage and accompanying control documents. Separately, we assemble a corpus of 127,518 state trial court complaints, which we use for analysis.
- We introduce ScoreMix, a detector-agnostic method that turns detector scores into a calibrated prevalence estimate (Section 4.1), and best practices for estimating AI usage in trial courts: we validate settings on text with known AI usage, and correct estimates with the measured error rates. Almost all prior work does neither.
- We apply ScoreMix to complaints from four state trial courts and find that (1) estimated AI usage rose to 35.6% of complaints in San Francisco and 37.8% in Santa Clara, compared with around 16% in the Oklahoma counties; (2) an estimated 71.6% of pro se complaints in Santa Clara show AI usage; and (3) detection is concentrated in two sections of the complaints, the statement of facts and the causes of action.

[1] https://huggingface.co/datasets/Arifov/PhantomDocket

2 Problem Setting and Background

Our primary corpus of study are civil complaints filed in four state trial courts: Oklahoma, San Francisco, Santa Clara, and Tulsa. We begin by discussing how we collect complaint data and then describe three existing AI detection methods: word-occurrence MLE, EditLens, and Pangram (Section 2.2).

2.1 Problem Setting: Civil Complaints in State Trial Courts

There is growing concern among judges that generative AI is entering courtrooms at the state trial court level, taking over resources, and overwhelming the justice system [44]. State courts receive the overwhelming majority of U.S. court filings, averaging 98.6% of all cases since 2012 [9], but public filing archives such as RECAP cover filings at the federal level ~~[9]~~[16]. Data access at the state court level faces multiple barriers, including fragmented systems, institutional restrictions, and poor data quality [41]. We address this data challenge by collecting a new dataset of filings from four state trial courts. We developed custom scrapers to collect filings from three digital systems available publicly, resulting in data from four trial courts, two in Oklahoma (Tulsa and Oklahoma counties) and two in California (San Francisco and Santa Clara counties). The detector corpus contains 31,143 case-initiating filings from Tulsa, 46,740 from Oklahoma, 25,134 from San Francisco, and 24,501 from Santa Clara filed from 2020–2026. Annual counts by jurisdiction and year are reported in Appendix A.1 (Table 4). We collect case-initiating pleadings from the following case types: civil cases seeking more than $10,000 in the Oklahoma counties; general civil cases in San Francisco County; and Santa Clara County civil actions involving tort, contract, employment, civil-rights, business, property, insurance, intellectual-property, malpractice, and other non-collection complaints. These criteria differ by court and limit cross-jurisdictional comparisons. All datasets, except for San Francisco, represent a census of every eligible pleading under the selection criteria. For comparative analyses below, we examine subsets of cases that are closest across jurisdictions.

Source filings were downloaded as PDFs or TIFF images and converted to text using OCR. We exclude Judicial Council forms, which offer streamlined checkboxes and write-in boxes for common claims.

2.2 Background: AI Detection and Existing Methods

As generative AI use has proliferated, so has work on detecting and measuring it [54]. Current AI-text detection approaches fall into four broad categories [54]: *statistical methods*, which observe properties of real-world text using reference language model (e.g. the entropy of logit distributions) [3, 18, 32]; *neural classifiers*, which learn from examples of human and AI-generated text [14, 48, 50]; *watermarking* instead embeds a detectable signal during generation [21] and human-based detection. No method is generalizable: benchmarks show that detectors performing well in-distribution are often brittle in practice [13, 19, 51] due to: text from unseen models [38, 55], paraphrasing, and adversarial rewrites [13, 23], distribution shift across linguistic domains, populations, and time periods [25, 36, 38]. One primary split in AI-detection is whether to study *document-level detection* or *population-level prevalence*. Population-level estimation methods [24, 26] tend to be cheaper to run but often perform worse on smaller samples and bias grows as human writing drifts from the reference period [38]. We study three methods: one corpus-level estimator, word-occurrence MLE [24]; and two document-level classifiers, EditLens [48] and Pangram 3.3.2 [30]. All three depend on synthetic AI text, either as training data or as a reference distribution. Throughout, we refer to P and Q as distributions of human-written and AI-generated text, respectively.

Word-Occurrence MLE. is a distributional method designed to estimate the prevalence of AI-like writing, α_{MLE}, in a corpus [24, 27]; we refer to it as the MLE. The method models text, x,

as coming from either a human distribution P or an AI distribution Q. α_{MLE} is estimated by directly maximizing the likelihood, $\hat{\alpha}_{\text{MLE}} = \arg\max_{\alpha_{\text{MLE}} \in [0,1]} \prod_i [(1 - \alpha_{\text{MLE}})P(x_i) + \alpha_{\text{MLE}} Q(x_i)]$. P is estimated using randomly selected human text (in our case, a random subset of filings from 2020–2022). The AI distribution Q is estimated using model rewrites for each text in the P fitting set. In other words, the model estimates Bernoulli word-occurrence probabilities under P and Q. For our purpose, we draw from Q by first converting source materials into notes and then generating the document based on notes (using GPT-4o-mini), holding length approximately fixed. We fit separate jurisdiction-specific reference distributions. Before estimation, we also take the following preprocessing steps: we tokenize and remove OCR artifacts, remove exact duplicate sentences appearing in the reference or target duplicate banks, and divide the remaining text into consecutive 40-token chunks. These preprocessing choices adapt the MLE estimator to the distinctive structure of legal filings (Appendix B.2). To obtain document-level quantities for comparison with Pangram and EditLens, we pool all 40-token chunks belonging to a document and fit one mixture MLE jointly to those chunks, yielding $\hat{\alpha}_{d,\text{MLE}}$. This document-level quantity is a substantially noisier application, as MLE is intended for corpus-level prevalence estimation. At the corpus-level, however, MLE is substantially less resource intensive and expensive than document-level models described next.

EditLens. is an open-weight transformer-based detector trained on large amounts of human-written texts and synthetically generated AI versions of this text[2] This model is directly trained, as a classifier, to separate P from Q [48]; its power rests on the data it is trained on. To identify AI generated text, authors generate text using AI editing instructions of varying intensity to capture a variety of use-cases and writing styles. Text is also edited a number of times; the number of times a text is edited is tracked as a sign of intensity and for each text input, EditLens produces four probabilities, p_k for $k = 0, \ldots, 3$, representing increasing likelihood of AI involvement, which we integrate over to generate a single score: $s = \frac{1}{3} \sum_{k=0}^{3} k\, p_k$. We divide each filing's raw OCR text into consecutive, nonoverlapping 400-word windows, and calculate a document's score as the unweighted arithmetic mean of its window scores (Appendix B.2). Because EditLens is an open model, it is substantially cheaper to run than proprietary methods.

Pangram. is a proprietary commercial detector that has drawn wide attention. Independent evaluations report strong accuracy [20] but also show that it can be evaded [6]; further, it is very expensive, at $0.05 per 1000 words for Pangram 3.3.2, making systematic assessments of large corpora increasingly infeasible. Pangram has been applied to legal writing in several recent studies: 1,600 federal civil complaints [45], roughly 2,250 federal appellate opinions [33], and 571 articles from the flagship law reviews of the T14 [34]. Prior work, however, does not measure sensitivity against legal text of known AI authorship, does not report precision and recall on legal writing, and relies on Pangram's model card or general-domain benchmarks for accuracy [20]. In this work, we use Pangram 3.3.2, a release of Pangram 3.3 [30], which uses architecture based on the EditLens model [48]. Like EditLens, Pangram is a trained classifier that separates human text from AI rewrites of human text [14], but its training data and weights are proprietary, so its Q can be neither inspected nor adapted; we can only calibrate its outputs (Section 4.1). The Pangram API returns outputs at both the document and segment levels, including a categorical document prediction (`Human`, `Mixed`, or `AI`) and a normalized AI-assistance score between zero and one for each segment (Appendix B.2). Pangram cautions that templated writing, bullet lists, technical manuals, tables of contents, and artifacts introduced by PDF parsing are especially susceptible to

[2]We use the released `pangram/editlens_Llama-3.2-3B` checkpoint, a LoRA adapter for Llama-3.2-3B, without further fine-tuning.

false positives [15, 30]. In our work, we construct a continuous document score as the unweighted arithmetic mean of the returned segment-level AI-assistance scores.

3 Why AI Detection in Legal Text Is Hard

AI detection is difficult in legal text and a detector's default settings cannot be trusted without validation. Legal writing contains a large share of boilerplate and uses a narrow vocabulary register, that a detector calibrated on general text can misread. Further, the detectors themselves require many hyperparameter settings, and prevalence estimates vary widely based on these. We discuss these challenges here, and Section 4 presents our approach to both.

3.1 Theoretical Limits: Legal Text Is Hard for Any Detector

The ideal performance of any AI-detector is bounded by the extent to which the AI text distribution Q differs from the human text distribution P: AUROC $\leq \frac{1}{2} + \mathrm{TV}(P,Q) - \mathrm{TV}(P,Q)^2/2$, where $\mathrm{TV}(P,Q)$ is the Total Variation distance between the distributions [42]. Several properties of a vocabulary can affect this, particularly in specialized domains (e.g. law); we discuss these now.

Challenge #1: Boilerplate content makes AI detection harder; legal text in our dataset is 26% boilerplate. We label sequences of 8 tokens as boilerplate if they appear in at least 1% of a domain's human documents; a word *anchored*, or *forced*, if it always appears in both human documents and AI regenerations of those documents (Appendix C.3). We find that, in legal text, about 26% of tokens are boilerplate and another 9% are anchored; in more *general-use* domains (e.g. press releases, scientific abstracts) at most 6% of units are (Table 14). See Appendix D for more details on how each of its columns is computed. If a share s of a unit is authored and the rest is shared, the bound becomes AUROC $\leq \frac{1}{2} + s\,TV(P,Q) - s^2 TV(P,Q)^2/2$ (Appendix[3] C.5). Averaged over units, the ceiling for legal text is 0.88 when the authored part is perfectly separable and 0.76 at $T = 0.5$, near the distance measured above (compared with 0.96–0.98 and 0.80–0.83 for the other domains). The effect follows the law: $\Phi^{-1}(\mathrm{AUROC})$ falls in proportion to the authored share ($\hat{\gamma} = 0.98$); a 40-token legal unit that is half boilerplate drops in separability from AUROC 0.986 to 0.870 (Appendix C.5). Prior research has found boilerplate also scores formulaic *human* text as AI (resulting in higher detection rates for non-native English writers [25]). Our explorations show that templated filings score as human under the word-occurrence detector (Appendix C.4), and EditLens scores are nearly uncorrelated with boilerplate share (Appendix C.7).

Our boilerplate measure is also a *lower bound*: it detects only text that repeats across filings, so a lease or account statement reproduced once goes uncounted even though the filer did not write it.

Challenge #2: Legal drafting is a narrow-vocabulary task. In a domain with more consistent writing patterns and fewer *choices* per word, it's likely that less of a model's distinct signature can show. We measured the model's own next-token entropy while it drafts each document (Llama-3.1-8B-Instruct, 200 documents per domain, five samples each). Drafting a petition runs at 0.35 nats per token (a measure of cross-entropy or surprise) compared with the more general domains studied in [26, 27]: 0.43 for a press release, 0.45 for a UN release, 0.53 for an abstract, and 0.57 for a consumer complaint (each difference significant at $p < 10^{-30}$ by Welch's t). When the register is restricted, model output will generally be more constrained. The rate of the word *plaintiff*, for instance, increases from 17 to 28 per 1,000 tokens; the word *court* increases from 2.8 to 11.6 per 1,000 tokens. For press releases, 44% of the distance is comprised of idiosyncratic words. In our setting, shared material bring P and Q closer together (Appendix C.9). A narrow register also harms

[3] If a unit's forced tokens come from a distribution B shared by both writers and its authored tokens from P_{auth} or Q_{auth}, then the mixtures $P = (1-s)B + sP_{\text{auth}}$ and $Q = (1-s)B + sQ_{\text{auth}}$ satisfy $\mathrm{TV}(P,Q) = s\,T$ exactly

a detector that was calibrated on a more general register: one of our detectors (Section 2) falls from document AUROC 0.93 on plain drafts of held-out petitions to 0.77 on register-prompted drafts of the same petitions,[4] against 0.93 to 0.83 in press releases (Appendix C.6).

Challenge #3: matching real-world AI use-cases is essential for high detector capability; how people are using legal AI is unknown. Many AI detectors in the literature [26] generate AI-text in order to estimate distributions, Q. The choice of how to generate AI-text influences the detector's performance on real-world texts. For example, in MLE's original implementation, authors [26] ask a model to `rewrite document: {text}`. We find that the Q that emerges from this process differs greatly from a Q' that emerges from other generation processes. For example, what would Q' look like if it was generated not from requests to rewrite text but from answers to user questions? We designed the following experiment to emulate this setting: we took 378 passages from human-written legal petitions; for each passage, we had agents write 1,134 questions a pro se filer or an attorney might plausibly have asked themselves that their passage answers; then, we had three open models answer *these* questions directly (Appendix C.8); from this set, we estimate Q'. We then demonstrate how detectors trained on P vs. Q (i.e. AI text generated with a `rewrite` prompt) differ from detectors trained on P vs. Q' (i.e. AI text generated in response to questions) on a held-out mixed set of AI-generated rewrites, AI-responses to questions, and human text. An AI detector trained on P and Q, we find, flags 90% of AI-generated rewrites but only 16–72% of AI-responses to questions (with spread due to three AI-generating models and three ways of phrasing questions). Conversely, a detector refit on Q' separates those same responses from human text at 89–100% over those same nine combinations. Thus, AI detectors can only be effective if they have a reasonable case for what the use-cases in the target population are. (Appendix C.8, Section 6).

These challenges are not unique to law. While AI-detection has shown promise, heavy boilerplate, a narrow register, and unknown patterns of AI usage will affect the accuracy of detectors in any professional domain.

3.2 Practical Challenges: Detectors Are Sensitive to Their Settings

Researchers typically have to make decisions regarding how the text is scored, *model hyperparameters*, and how the scores are aggregated, *aggregation rules*. Both ultimately shape the outcome of AI detection. We define *settings* as some combinations of model hyperparameters and aggregation rules (Table 1). Across attempted combinations of EditLens settings, the share of filings flagged as AI-generated ranged from under 5% to almost 50% (Appendix E.6). While black-box detectors like Pangram lack tunable hyperparameters, they too are subject to aggregation variation. Across nine reasonable Pangram document-score constructions, the share of filings classified as positive ranged from 22.3% to 52.4% (Appendix E.5). Even after choosing an aggregation rule, a classification threshold is still needed to separate classes.[5] On a selection of complaints from Santa Clara, classifying a document as AI if the mean EditLens window score is at least `0.10` flags 9.7% of filings, `0.05` flags 33.7%, `0.01` flags 97.5% (Table 21).

Prior work on court filings typically does not acknowledge such variance: [33, 34, 45] all report one setting and typically validate it only against pre-ChatGPT filings, which bounds false positives but not false negatives (filings from before ChatGPT are assumed all human-written). But without filings *known* to contain AI text, there is no way to measure how many AI-assisted filings the setting misses: a setting that flags nothing before November, 2023 may also flag very little after it.

[4] Appendix C.6 gives the prompts and process we use to generate drafts.
[5] To compare settings, we selected a separate threshold for each Pangram and EditLens setting using Youden's index. E.5 E.6

Method	Setting	Options	Used
Pangram	text submitted	2	OCRed text
	document score	8	mean window `ai_assistance_score`
EditLens	window length	18	250 words
	stride fraction	7	0.60
	terminal-window rule	3	retain all
	window signal	8	$1 - p_0$
	document score	16	mean
MLE	estimation unit	3	40-token chunks
	vocabulary cutoffs	10	(5, 3), no POS filter
	human reference P	2	same court, 2020–22
	AI reference Q	5	LLM rewrite of redacted skeleton
	preprocessing	3	processed text + sentence dedup.
	document score	5	pooled corpus-level $\widehat{\alpha}_{\text{MLE}}$

Table 1. Number of values considered for each detector setting, and the value we use. Rows colored green are *aggregation rules*, the choices made *after* inference about how a document is labeled as AI-assisted; rows colored purple are *model hyperparameters*, which change the scores the model produces. Table 15 (App. E) lists every value.

4 Our Approach: Validating and Calibrating Detectors on PhantomDocket

Detectors should be validated and calibrated on known legal text with known AI usage. We first present our method, ScoreMix, which turns any detector's scores into a calibrated estimate of AI prevalence (Section 4.1). We then introduce PhantomDocket, the benchmark of real court filings with documented AI (Section 4.2). Finally, we validate the detectors, their settings, and the calibration on PhantomDocket (Section 4.3).

4.1 Our Method: Turning Detector Scores into Calibrated Prevalence

We introduce ScoreMix, a method that turns the scores of any AI detector into a calibrated estimate of the share of filings whose detector scores resemble those of the hallucination filings. Let s_i be the detector's score for document i. Following the same structure as the MLE mixture, ScoreMix models s_i as drawn from the mixture $(1-\alpha)\,P_S(s_i) + \alpha\,Q_S(s_i)$, where P_S is the density of scores on human-written filings, Q_S is the density on filings with AI usage, and α is the share of filings drawn from Q_S. Both densities are estimated from documents with known AI usage: P_S from the court's own filings before the release of ChatGPT, and Q_S from the hallucination filings in PhantomDocket (Section 4.2). α is then estimated by maximum likelihood with the EM algorithm with a bootstrap confidence interval (Appendix F). This is mathematically equivalent to the probabilistic classify-and-count methods that adjust imperfect detectors [4, 5, 43].

An alternative to ScoreMix is the well-known Rogan–Gladen estimator [39]: $\hat{\alpha}_{\text{RG}} = (\hat{q} + \text{spec} - 1)/(\text{sens} + \text{spec} - 1)$, where $\hat{q}$ is the share of filings with score above a threshold, sensitivity is measured on the hallucination filings in PhantomDocket, and specificity on the court's own 2020–2022 filings. The Rogan-Gladen estimator is used as a natural comparison to ScoreMix in Section 4.3, but requires a choice in threshold which introduces sensitivity.

We chose to apply ScoreMix on EditLens because it (a) performed well on PhantomDocket, (b) is cost-effective to run as an open model, and (c) is easy to tune to our task. This assumes that AI usage in the trial courts produces similar EditLens scores to the AI usage in PhantomDocket. Q_S can also be estimated from EditLens scores on the synthetic generations (Section 2.2).

We note that α is an estimated mixture weight on the EditLens score distribution observed on the PhantomDocket hallucination filings. It estimates the share of complaint filings whose scores arise from that distribution; it is not the rate at which filings contain a hallucination. Additionally,

positive documents in PhantomDocket are mostly briefs and motions, whereas this corpus consists of complaints, and these differ in register and boilerplate. Section 4.3 reports how much of the dependence on detector settings this calibration removes.

4.2 Constructing PhantomDocket, Our Ground-Truth Dataset

A primary challenge to trustworthy AI detection in the legal setting is the lack of ground-truth data: as Section 3 showed, synthetic AI text may not represent how AI is actually used in drafting [22], and no prior work on court filings has validated a detector on legal writing.

To address this challenge, we look towards known, demonstrable examples of AI-usage in court filings: Charlotin's AI Hallucination Cases Database (Charlotin Database) [7] catalogs legal decisions where a judge discusses a litigant's use of an AI generated phantom citation or hallucinated argument in their filings. These examples, we reason, are as close to ground truth as exists without fully observing human lawyer drafting processes. While Charlotin tracks the PDF decisions and case metadata, the underlying filings with implicated hallucinations themselves are not systematically available, as it requires extensive docket searches and costly retrievals (e.g., purchasing federal filings from PACER when they are not in the CourtListener RECAP archive) to identify these. We introduce the first dataset linking these decisions to the docketed hallucinated filings, called PhantomDocket. This dataset differs from previous works on legal hallucinations, which evaluate hallucination rates [11, 29] or how well models detect synthetic hallucinations injected into human-drafted briefs [28].

To create PhantomDocket, we first filtered down to deduplicated U.S. cases in the Charlotin Database. For each case, we searched for the filing that includes the hallucination using these steps: (1) Extraction of case docket number from the ruling document's caption or document header. (2) Identifying the electronic filing number (ECF) or docket-entry number, filing title, and filing party of the implicated document. (3) Locating the case in Lexis CourtLink or Bloomberg Law and matching the candidate to an exact docket entry using the docket number, court, case caption, filing number, and filing party. (4) Retrieving the filing and verifying its identity using the internal CM/ECF stamp, caption, and title. If available, we also verified that the target filing includes the hallucination quoted in the ruling. This yields 980 hallucination filings drawn from 619 cases.

This process required resolving docket numbers that are unavailable in the database, mapping rulings that do not name the offending filing to a specific docket entry, and retrieving filings that are not freely available. We describe these challenges, and our verification of each retrieved document, in Appendix A.2. In our dataset, many litigants expressly acknowledge that hallucinations were introduced through AI use. That said, we cannot observe the extent and precise nature of AI utilization for each document. While the Charlotin data provides the most comprehensive starting point, it can introduce selection bias into our sample. AI may be utilized without resulting in a hallucination, and other data collection protocols were not fruitful at collecting such documents. If our observed instances represent less careful uses of AI, it is possible that AI use estimates tuned to this benchmark will provide conservative estimates of AI use. That said, PhantomDocket thereby offers the first systematic dataset of full filing documents known to have used AI, and the set of methods we examine are AI detection methods, not hallucination detection methods. We nonetheless also supplement this sample with documents drafted by the same party implicated in using AI, but ones not involving hallucinations.

While the documents described above are the best sample of legal filings with verified use of AI, we also require negative sample that controls for case characteristics for benchmarking. We hence also include two other sets of documents from the same cases: filings by opposing counsel and judicial opinions. To restrict leakage of AI text into these document sets, we selected documents which appeared before the offending filing in the docket. We assume AI was not used to the same

Detector	AUROC (95% CI)	Detector	AUROC (95% CI)
Pangram 3.3.2	0.946 [0.934, 0.958]	MLE, Santa Clara P, Q fit	0.767 [0.746, 0.788]
EditLens	0.896 [0.881, 0.911]	MLE, San Francisco P, Q fit	0.761 [0.739, 0.783]
MLE, Oklahoma P, Q fit	0.777 [0.754, 0.798]		

Table 2. **Detector performance on PhantomDocket.** AUROC on the common evaluation set, separating 953[7] filings with documented AI hallucinations from 771 controls (509 opposing-party filings and 262 judicial orders), with 95% confidence intervals from document-stratified bootstrap draws. Each detector's setting is described in Table 1. Each MLE score is a document-level mixture estimate $\hat{\alpha}_{d,\mathrm{MLE}}$ with P and Q fit on the indicated jurisdiction.

extent in the drafting of opinions and opposing counsel filings This design of PhantomDocket is crucial for validation. Given that all of these document sets are drawn from the same case dockets, a detector is not confounded by jurisdiction, year, area of law, or the register of a particular judge's chambers; what separates them are the properties of the filings themselves. The full inventory funnel (Charlotin source → collected documents) is in Appendix A.3, Figures 6 and 7.

4.3 Validation on PhantomDocket

We use PhantomDocket to (1) compare how well each detector separates AI from human filings, (2) choose the settings of the detector we use, and (3) measure how much the calibration in Section 4.1 reduces the dependence of our estimates on those settings.[6] For each of the three PhantomDocket sets (hallucination, same party and control), we ran each of the documents through Pangram, EditLens and three versions of MLE, with P and Q fit on the Oklahoma, Santa Clara, or San Francisco complaints (13). Table 2 shows how well each method can classify documents as $y = 1$ (i.e. containing AI-generations, proxied by whether or not it contains a hallucination) or $y = 0$. As can be seen, all five detectors rank hallucination filings above controls, but not equally. Pangram reaches an AUROC of 0.945, EditLens 0.896; the three MLE fits cluster around ≈ 0.77.

We use PhantomDocket to choose among different EditLens settings in Table 1. A Tree-Structured Parzen Estimator (TPE) hyperparameter tuning procedure (Appendix G) selected the optimal setting to consist of a 250-word window, stride of 0.60, window scores using the probability assigned to any-AI bucket ($1 - p_0$), and averaging across windows. This reached a partial AUROC of 0.841 on a locked test set. We contrast it with the simple setting we began with, 400-word windows without overlap scored by the expected bucket and averaged, which reached a partial AUROC of 0.785; the paired difference was 0.056 (95% CI 0.017–0.093). Turning either setting into a document-level label requires an operating threshold. For each court, we set thresholds that flag 1%, 5%, and 10% of its 2020–2021 human-written complaints. The proportion of detected complaints in 2026 varied by 18–22 percentage points across the settings and thresholds. This reduced to 12–17 points when we calibrated with Rogan–Gladen and used recall on PhantomDocket and specificity for each court.

Each threshold discards information about where each score falls. Using filings from PhantomDocket and applying ScoreMix allows us to use the full score instead. Under the two settings, its 2026 estimates differed by 0.6–5.6 percentage points across courts, which is smaller than the range using Rogan–Gladen and comparable to the widths of the selected setting's 95% bootstrap intervals (2.7–6.5 percentage points). In an analysis of 2022 complaints, ScoreMix estimated at most 0.5% in every court under either setting. Thus, ScoreMix reduces, but does not eliminate, the

[6] See Appendices E.4, **??** and E.7 for full results.

[7] MLE could not score 27 hallucination filings after preprocessing, reducing the common detector set from 980 to 953.

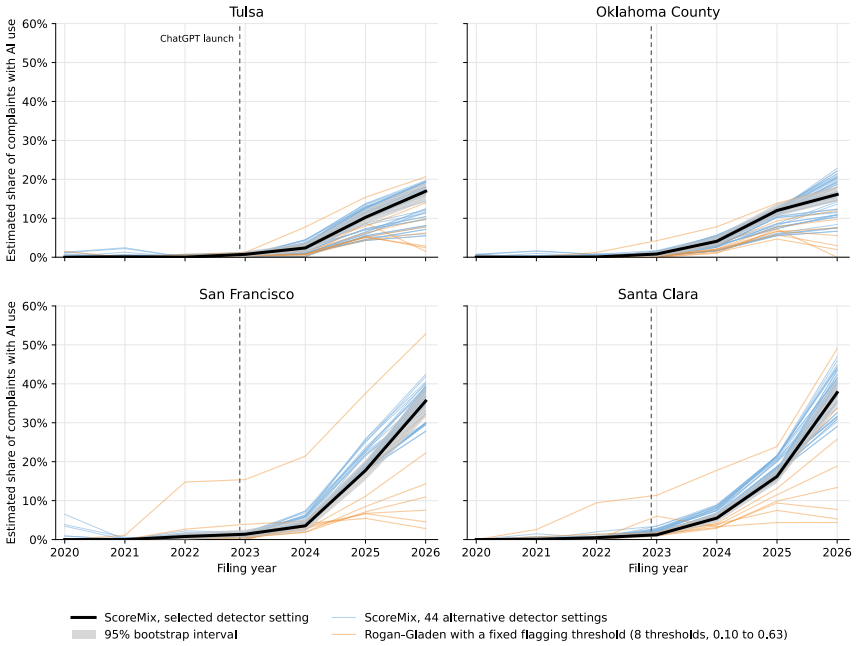


Fig. 2. **AI Prevalence estimates across courts, with variation shown across model settings.** We show the estimated share of a court's complaints with AI usage by filing year, as well as the variation that can arise under three families of estimates. The thick black line is ScoreMix on validated model settings, validated on PhantomDocket (i.e. 250-word windows, a 0.60 stride fraction, any-AI probability, and mean pooling). The thin blue lines are ScoreMix on 44 alternative detector settings (five window geometries, three window-score definitions and three pooling rules, with the final-window rule fixed at `retain`); their spread is the choice of detector setting alone. Thin orange lines are Rogan–Gladen estimates (Section 4.1) for eight thresholds from 0.10 to 0.63, using sensitivity from the PhantomDocket hallucination filings and the false-positive rate from the court's own 2020–2022 complaints. We demonstrated that, in 2026, the San Francisco estimate shifts from between 3% and 53% and the Santa Clara estimate between 4% and 49%, whereas the detector setting moves the ScoreMix estimate between 28% and 42% and between 29% and 47%. We exclude all collections/debt cases from the estimates in the figures.

dependence of prevalence estimates on different settings. We use the optimized EditLens setting with ScoreMix for the prevalence results that follow. We use the non-overlapping setting for the analysis in Section 5.3. Figure 2 shows both sources of variation, the threshold and the detector setting, for every court and year. See Appendix G.1 for more settings and aggregation rules.

We combine a competent detector that is cheap enough to run at scale on every filing in a court, a benchmark that measures how that detector and each of its settings perform on real legal AI usage, and a calibrated prevalence estimator for our following results.

5 Estimating AI Prevalence in Complaints

Having validated the detector's settings and its calibration, we now estimate the prevalence of AI usage in the complaints.

5.1 AI Prevalence Across Jurisdiction and Time

We see a dramatic rise in AI-like prevalence across all of our jurisdictions from 2024 to 2025, more than a year after the release of ChatGPT, ranging from 7.87 to 14.24 percentage point increases. By 2026, we measure $\hat{\alpha}^{\text{SC},2026} = 0.378$ (95% CI 0.344–0.412), $\hat{\alpha}^{\text{SF},2026} = 0.356$ (95% CI 0.315–0.391), $\hat{\alpha}^{\text{Tulsa},2026} = 0.169$ (95% CI 0.144–0.192), and $\hat{\alpha}^{\text{OK},2026} = 0.161$ (95% CI 0.143–0.181).

The reported absolute prevalence estimates are still sensitive to the choice of detector specification, which reinforces the importance of reporting alternative specifications alongside the selected setting. Figure 2 shows what the estimates look like with different thresholding and Rogan–Gladen corrections. The choice of threshold alone shifts AI-usage estimates from 2–21% in Tulsa, 0–18% in Oklahoma County, 3–53% in San Francisco and 4–49% in Santa Clara. The choice of detector under ScoreMix moves it across 6–20%, 7–21%, 28–42% and 29–47%.

Why is estimated prevalence higher in the California counties than in Oklahoma? Case mix offers some explanation. Appendix Table 28 reports each legal issue family's share of 2026 complaints and its corresponding ScoreMix prevalence estimate. Collections and debt matters constitute approximately 46–48% of filings in Oklahoma County, Tulsa, and San Francisco, but they were not collected in Santa Clara, between 0.0% and 1.2%. By contrast, some issue families have substantially higher estimates and happen to be concentrated in our California corpuses. Business and commercial matters reach 83.4% in San Francisco and 81.5% in Santa Clara, while employment matters reach 98.3% and 59.2%. Personal injury and tort complaints are comparatively similar across courts, ranging from 21.4% to 30.2%; however, contract estimates range from 8.5% to 21.1%. Case type composition plausibly accounts for some of the aggregate differences, in particular because collections constitute nearly half of the filings of some courts and estimate at zero prevalence. Because Santa Clara collections cases are structurally undercollected, its all-filings estimate is not fully comparable with those of other courts, which is why we exclude all collections and debt complaints from Figure 2.

We suspect that many court cases with less AI usage are more likely to be filed from templates, where there is little text to write; debt and collections complaints are often formulaic with very little narrative. On the other hand, in San Francisco's unlimited-civil docket categories are weighted toward employment and privacy, more variation requires complaints to be drafted individually. Industry surveys report broad adoption across firm types: 71% of solo practitioners, 75% of small firms and 86% of mid-sized firms say they have adopted AI [37].

5.2 *Who* is AI Enabling the Most? Comparing Pro Se Litigants vs. Attorneys and Examining Power Users

In Figure 3, AI prevalence rose much more quickly among pro se plaintiffs (those representing themselves) than those represented by an attorney in Santa Clara. In 2024, the estimate was 18.8% (95% CI 12.1–26.5; $n = 167$) for pro se complaints and 4.7% (95% CI 3.6–6.0; $n = 4{,}066$) for attorney-represented complaints. By 2026, the corresponding estimates were 71.6% (95% CI 63.6–79.1; $n = 266$) and 33.5% (95% CI 30.3–36.9; $n = 2{,}913$). Santa Clara was the only jurisdiction with at least 100 pro se complaints labeled I.2, satisfying our ScoreMix criterion.

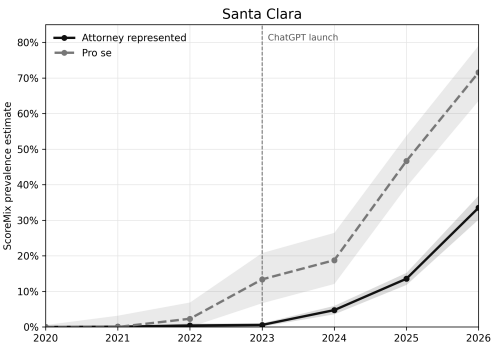


Fig. 3. AI prevalence across pro se vs attorney represented plaintiffs in Santa Clara.

Do the same attorneys keep using AI? Plaintiff attorneys are named on 92–98% of complaints,

Section of case	% Windows	Pr(higher)
Causes of action	58.5%	68.9% ± 1.4
Statement of facts	33.2%	65.7% ± 1.2
Definitions	0.5%	63.4% ± 4.7
Prayer for relief	18.6%	47.2% ± 1.3
Jurisdiction & venue	4.6%	47.6% ± 2.0
Verification, signature	7.4%	36.5% ± 2.9
Attachments & exhibits	55.1%	30.6% ± 1.2
Caption & parties	10.6%	46.2% ± 1.4
Venue	1.4%	38.4% ± 3.3

(a) Sections of a complaint; 2026.

Case type	% Complaints	Pr(higher)
Wrongful termination	6.5%	82.3% ± 1.7
Business tort	2.3%	78.2% ± 3.5
Product liability	1.2%	66.5% ± 5.1
Fraud	1.2%	68.8% ± 5.6
Professional malpractice	0.6%	67.9% ± 7.6
Contract/warranty	18.7%	47.9% ± 1.5
Auto	7.7%	44.4% ± 2.2
Insurance/bad faith	0.7%	31.4% ± 7.8
Debt collection	27.0%	18.7% ± 1.0

(b) Case types, California; 2026.

Table 3. **Sections of complaints and case types with highest EditLens scores.** Pr(higher) indicates the probability that a random passage in the category exhibits a higher AI score than the comparison. Pr(higher) of 50% indicates no difference. Scores were calculated using EditLens with 400-word, nonoverlapping windows. Panel (a) displays results for complaint sections. Panel (b) compares arithmetic-mean document scores across California case types and retains the five high-contrast and four low-contrast categories selected in the original mean-score analysis among types with at least 50 complaints. Values are estimates ± the larger half-width, in percentage points, of their 95% intervals from 5,000 jurisdiction-stratified bootstrap draws; panel (a) resamples document clusters.

and most file more than once, so we can track them over time. We look at attorneys with at least two filings in each of two consecutive years. If an attorney had at least one filing flagged in a given year (mean EditLens score of at least 0.10), the chance they are flagged again the next year is 27% in Oklahoma County, 35% in Tulsa, 55% in San Francisco, and 65% in Santa Clara. For attorneys who were not flagged, it is 5–8% across all jurisdictions. Among attorneys first flagged in 2024, 22–80% were flagged again in 2025 or 2026. This is not just an attorney's writing style being picked up by the detector: before ChatGPT, almost no attorney was flagged at this threshold, so there was nothing to persist. (At the lower 0.05 threshold we do see similar persistence before 2023, which is why we use the stricter one.) AI usage is also not concentrated among the biggest filers. The top 10% of attorneys by 2026 volume file 42–71% of complaints but account for only 33–43% of flagged ones, consistent with high-volume collections practices filing from templates.

5.3 What *parts* of a Complaint are Most Likely to Contain AI?

We report *where* our detectors believe AI is used in legal drafting. This analysis uses non-overlapping 400-word EditLens windows, and we label each of the 47,633 windows by its function, which are then mapped onto the sections of a civil complaint (Appendix I.3). Table 3(a) reports section's probability of superiority: the probability that a randomly selected window of that type has a higher detection score than a randomly selected window of other types, with half credit for ties. A value of 50% indicates no separation. We observe higher rates of AI usage in narrative sections, as opposed to more formulaic sections of complaints. Causes of action and statements of fact have a probability of scoring higher for AI utilization of 68.9% and 65.7%, respectively. In contrast, attachments and exhibits and verification and signature sections have probabilities of 30.6% and 36.5%, indicating lower propensity for AI utilization. These findings are broadly consistent with the challenge we highlighted in Section 3, namely that boilerplate can attenuate detection.

At the same time, the high prevalence in narrative sections is not solely due to the narrow register. Between 2020–21 and 2024–26 the probability of superiority rises for statements of facts (+11pp) and causes of action (+12.1pp) parts of the complaint (Table 33); (Appendix I.3).

6 Discussion

Detecting AI writing in professional domains like law is challenging due to narrow-vocabulary tasks, uncertainty about how AI is used in the writing process, and uncertainty around the validity of detectors OOD. For these reasons, we believe that running AI detectors on legal documents to draw document-level inferences should be done cautiously.

In this work, we leveraged the Charlotin database to develop PhantomDocket, which allows us to evaluate AI detectors on legal documents with plausible ground truth. We found that AI detectors perform well but not perfectly, with Pangram performing the best, but at an unscalable cost. Methods relying only on word frequencies performed the worst, and open weight models provided a cost-effective alternative for population estimation. While ScoreMix grounds our estimate in actual legal writing and AI usage, if the AI usage in the trial courts does not resemble the hallucinations in PhantomDocket, we expect to introduce some bias in detection. We also observe that estimates can increase dramatically after applying the ScoreMix and, ideally, we would be able to ground detection in known uses of AI.

Across four counties, we find substantial geographic variation in estimated AI-assisted drafting prevalence. In 2026, estimates in the California counties were roughly double those in the Oklahoma counties (Section 5.1). All four counties rank high nationally by AI user share. Oklahoma County and Tulsa County both fall in the top 15% of U.S. counties, and San Francisco and Santa Clara both fall in the top 5% [31]. These county-level measures; however, may not be transferable to the legal sector, where adoption may follow different geographic patterns. Proximity to Silicon Valley may have led to an earlier adoption among San Francisco and Santa Clara lawyers and litigants, but our results cannot yet identify geography as the cause of the difference. Case composition and collected corpus coverage further complicates our comparison, as our issue-family analysis suggests that both the types of cases filed and the differences across them affect the estimates (Section 5.1; Appendix I.1).

Studies have documented the rise in pro se litigants across U.S. courts at the federal level [8, 45]. In combination with our Santa Clara pro se estimate (Section 5.2), this may suggest that AI has indeed lowered the barrier to file a case by a pro se litigant, but it does not yet answer questions of representation quality. We leave for future study whether these complaints contain hallucinations or whether AI use by pro se litigants impacts judicial outcomes.

Misleading AI detection can have serious disciplinary consequences and biased outcomes on certain populations [25]. Indeed, our benchmark shows that even with the best performing detectors, the false negative and false positive rates can be nontrivial. Moreover, AI detection in itself will only become harder as language models are trained to mimic human text and style [42]. Judges, attorneys, and researchers should interpret document-level results with an abundance of caution.

The disclosure of AI utilization may be valuable to understand adoption patterns across the courts, particularly when disclosures expose the particular process with which litigants used AI. A patchwork of standing orders across judges already imposes various degrees of requirements, from AI disclosures to strict prohibitions of AI use [40]. Policymakers, too, may propose disclosure amendments to codes of civil procedure, as they have done so in California [17]. As it stands, however, work like PhantomDocket can help us calibrate our uncertainties and study AI prevalence in the courts. The sharp rise we document here corroborates that AI in the law is a question of how, not whether. And the first step to managing its risks is to understand its prevalence.

7 Conclusion

We have sought to bring empirical evidence and methods to an important issue in the legal community around the rise of AI in court filings. Detecting AI usage in this context is challenging

and current methods should be benchmarked with ground truth datasets. This is important both for assessing performance at an individual document level and to be able to calibrate the scores when applied to new sets of documents. We hope the findings inform understanding around the rise of AI in law and the methodological conversations around AI detection.

Acknowledgments

We are grateful to helpful comments from Jane Castleman, Lindsey Gailmard, Chris Manning, Derek Ouyang, Mirac Suzgun, and RegLab roundtable participants, the Stanford Institute for Human-Centered Artificial Intelligence (HAI) and Stanford Impact Labs for research support, and Damien Charlotin for collecting, maintaining, and making the AI Hallucination Cases Database publicly available. Pangram Labs provided research credits for the Pangram API and granted access to the gated EditLens model (`pangram/editlens_Llama-3.2-3B`), but had no role in the study design, analysis, or writing of this paper.

AI Disclosure

We relied on AI agents to develop our dataset collection tools and analysis scripts. We also used generative AI to review, suggest edits to this draft, and assist in compiling appendix materials.

References

[1] Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. 2020. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *International conference on machine learning*. PMLR, 222–232.

[2] American Bar Association. 2024. 2024 Artificial Intelligence TechReport. https://www.americanbar.org/groups/law_practice/resources/tech-report/2024/2024-artificial-intelligence-techreport/. Accessed: 2026-09-23.

[3] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. 2024. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *International Conference on Learning Representations (ICLR)*.

[4] Antonio Bella, Cesar Ferri, José Hernández-Orallo, and Maria Jose Ramirez-Quintana. 2010. Quantification via probability estimators. In *2010 IEEE International conference on data mining*. IEEE, 737–742.

[5] Dallas Card and Noah A Smith. 2018. The importance of calibration for estimating proportions from annotations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1636–1646.

[6] Tuhin Chakrabarty, Jane C. Ginsburg, and Paramveer Dhillon. 2025. Readers Prefer Outputs of AI Trained on Copyrighted Books over Expert Human Writers. *arXiv preprint arXiv:2510.13939* (2025).

[7] Damien Charlotin. 2025. AI Hallucination Cases Database. https://www.damiencharlotin.com/hallucinations/. Accessed 2026-09.

[8] Or Cohen-Sasson. 2026. The New Pro Se: Generative AI and the Surge in Federal Civil Self-Representation. *arXiv preprint arXiv:2605.29493* (2026).

[9] Court Statistics Project. [n. d.]. Federal and State Court Caseloads: Trends, 2012–2024. National Center for State Courts. State vs federal case loads. Interactive dashboard. Accessed 2026-10-07. https://www.ncsc.org/explore-court-caseload-data

[10] James H Curlin IV. 2025. ChatGPT Didn't Write This... or Did It? The Emergence of Generative AI in the Legal Field and Lessons from Mata v. Avianca. *Ark. L. Rev.* 78 (2025), 123.

[11] Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E. Ho. 2024. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis* 16, 1 (2024), 64–93. doi:10.1093/jla/laae003

[12] Jonas Dolezal, Sawood Alam, Mark Graham, and Maty Bohacek. 2026. The Impact of AI-Generated Text on the Internet. *arXiv preprint arXiv:2604.26965* (2026).

[13] Liam Dugan, Alyssa Hwang, Filip Trhlík, Josh Magnus Ludan, Andrew Zhu, Hainiu Xu, Daphne Ippolito, and Chris Callison-Burch. 2024. RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. 12463–12492. doi:10.18653/v1/2024.acl-long.674

[14] Bradley Emi and Max Spero. 2024. Technical Report on the Pangram AI-Generated Text Classifier. *arXiv preprint arXiv:2402.14873* (2024).

[15] Bradley Emi, Katherine Thai, Elyas Masrour, and Max Spero. 2026. Pangram 3.1 Model Card. Pangram Labs. Accessed 2026-10-12. https://www.pangram.com/research/model-card/pangram-3-1

[16] Free Law Project. 2026. RECAP Archive. https://www.courtlistener.com/recap/. Data file. Accessed 2026-10-14.

[17] Howard S. Glazer, Amy Jane Longo, Shannon Capone Kirk, Jafar Khalfani-Bey, and Britteny Leyva. 2026. California Enacts SB 574, Nation's First Law Governing Attorneys' Use of Generative Artificial Intelligence. Ropes & Gray LLP Insights. Alert published October 7, 2026. Accessed: 2026-10-17. https://www.ropesgray.com/en/insights/alerts/2026/10/california-enacts-sb-574-nations-first-law-governing-attorneys-use-of-generative-artificial

[18] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 17519–17537.

[19] Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. MGTBench: Benchmarking Machine-Generated Text Detection. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2251–2265. doi:10.1145/3658644.3670344

[20] Brian Jabarian and Alex Imas. 2025. *Artificial Writing and Automated Detection.* Working Paper 2025-116. Becker Friedman Institute for Economics, University of Chicago. Also NBER Working Paper 34223, September 2025. https://bfi.uchicago.edu/wp-content/uploads/2025/09/BFI_WP_2025-116.pdf

[21] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. 17061–17084.

[22] Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. 2025. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances* 11, 27 (2025), eadt3813.

[23] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense. In *Advances in Neural Information Processing Systems*, Vol. 36. 27469–27500.

[24] Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. 2024. Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. 29575–29620. arXiv:2403.07183.

[25] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4, 7 (2023). doi:10.1016/j.patter.2023.100779

[26] Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. The widespread adoption of large language model-assisted writing across society. *Patterns* 6, 12 (2025), 101366. doi:10.1016/j.patter.2025.101366

[27] Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Y. Zou. 2025. Quantifying large language model usage in scientific papers. *Nature Human Behaviour* 9, 12 (2025), 2599–2609. doi:10.1038/s41562-025-02273-8

[28] Patty Liu, Dominik Stammbach, and Peter Henderson. 2026. Who Checks the Citations? Benchmarking Legal Hallucination Detection. *arXiv preprint arXiv:2606.21155* (2026).

[29] Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning, and Daniel E. Ho. 2025. Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. *Journal of Empirical Legal Studies* 22, 2 (2025), 216–242. doi:10.1111/jels.12413

[30] Elyas Masrour, Max Spero, Bradley Emi, and Katherine Thai. 2026. Pangram 3.3 Model Card. Pangram Labs. Accessed 2026-10-12. https://www.pangram.com/research/model-card/pangram-3-3

[31] Amit Misra, Jane Wang, Will Fein, Scott McCullers, Kevin White, and Juan Lavista Ferres. 2026. *Measuring AI Diffusion Across U.S. Geographies: County, State, and Metro Estimates.* Technical Report MSR-TR-2026-18. Microsoft AI for Good Lab. https://www.microsoft.com/en-us/research/publication/measuring-ai-diffusion-across-u-s-geographies-county-state-and-metro-estimates/

[32] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 24950–24962.

[33] Josh Morrow. 2026. Signs of AI Authorship in Federal Appellate Opinions. The Volokh Conspiracy (Reason). Accessed 2026-10-12. https://reason.com/volokh/2026/08/24/signs-of-ai-authorship-in-federal-appellate-opinions/

[34] Julian Nyarko. 2026. How Often Are Articles in Top Law Reviews Written in Part by AI? The Volokh Conspiracy (Reason). https://reason.com/volokh/2026/09/07/how-often-are-articles-in-top-law-reviews-written-in-part-by-ai/.

[35] Riana Pfefferkorn. 2025. Who's Submitting AI-Tainted Filings in Court? Stanford Center for Internet and Society blog. Accessed 2026-11-18. https://cyberlaw.stanford.edu/blog/2025/10/whos-submitting-ai-tainted-filings-in-court/

[36] Shushanta Pudasaini, Luis Miralles-Pechuán, David Lillis, and Marisa Llorens Salvador. 2026. Why AI-Generated Text Detection Fails: Evidence from Explainable AI Beyond Benchmark Accuracy. *arXiv preprint arXiv:2603.23146* (2026).

[37] Catherine Sanders Reach. 2026. By the Numbers: What Surveys Show About Law Firm AI Adoption. North Carolina Lawyer, North Carolina Bar Association. Summarising the 2026 Clio, Thomson Reuters and 8am surveys of law firm AI adoption. Accessed 2026-10-14. https://www.ncbar.org/nc-lawyer/2026-05/by-the-numbers-what-surveys-show-about-law-firm-ai-adoption/

[38] Kevin Ren, Manish Raghavan, and Nikhil Garg. 2026. Hitting a Moving Target: Test-Time Adaptation for AI Text Detection under Continual Distribution Shift. In *Advances in Neural Information Processing Systems (NeurIPS)*. arXiv:2606.25152.

[39] Walter J. Rogan and Beth Gladen. 1978. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology* 107, 1 (1978), 71–76.

[40] Ropes & Gray LLP. [n. d.]. Standing Orders, Local Rules, and Decisions on the Use of AI. Artificial Intelligence Court Order Tracker. Accessed: 2026-10-14. https://www.ropesgray.com/en/sites/artificial-intelligence-court-order-tracker

[41] Tanina Rostain. 2024. Access to Justice as Access to Data. *Northwestern University Law Review* 119, 1 (2024), 5–22.

[42] Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2025. Can AI-Generated Text be Reliably Detected? Stress Testing AI Text Detectors Under Various Attacks. *Transactions on Machine Learning Research* (2025). arXiv:2303.11156.

[43] Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural computation* 14, 1 (2002), 21–41.

[44] Mattathias Schwartz and Zach Montague. 2026. Artificial Intelligence Floods Court Dockets with Home-Brewed Lawsuits. *The New York Times* (25 May 2026). Accessed 2026-10-07. https://www.nytimes.com/2026/05/25/us/politics/artificial-intelliegence-courts.html

[45] Anand Shah and Joshua Levy. 2026. Access to Justice in the Age of AI: Evidence from U.S. Federal Courts. SSRN working paper 6766859. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6766859

[46] Karen Sloan. 2026. Nearly Half of Early Career Lawyers Use AI Daily, Survey Finds. Reuters. Accessed: 2026-10-20. https://www.reuters.com/legal/legalindustry/nearly-half-early-career-lawyers-use-ai-daily-survey-finds-2026-07-07/

[47] James Taranto. 2026. The 'AI Detector' as Defamation Machine. The Wall Street Journal. Accessed: 2026-09-23. https://www.wsj.com/opinion/the-ai-detector-as-defamation-machine-8ba298f0

[48] Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. 2026. EditLens: Quantifying the Extent of AI Editing in Text. In *International Conference on Learning Representations (ICLR)*.

[49] Thomson Reuters Institute. 2026. *2026 AI in Professional Services Report.* Technical Report. Thomson Reuters. Accessed: 2026-09-23. https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/2026-ai-in-professional-services-report.pdf

[50] Vivek Verma, Eve Fleisig, Nicholas Tomlin, and Dan Klein. 2024. Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1702–1717.

[51] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, et al. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1369–1407.

[52] Benjamin Weiser. 2023. Here's What Happens When Your Lawyer Uses ChatGPT. The New York Times. Accessed: 2026-09-23. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

[53] Matteo Wong. 2026. America Has a Pangram Problem. The Atlantic. Accessed 2026-10-12. https://www.theatlantic.com/technology/2026/05/pangram-ai-detection-accuracy/687381/

[54] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *Computational Linguistics* 51, 1 (2025), 275–338. doi:10.1162/coli_a_00549

[55] Xiao Yu, Yi Yu, Dongrui Liu, Kejiang Chen, Weiming Zhang, Nenghai Yu, and Jing Shao. 2025. EvoBench: Towards Real-world LLM-Generated Text Detection Benchmarking for Evolving Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2025*. 14605–14620. doi:10.18653/v1/2025.findings-acl.754