Generative AI for Complex Document Review

A Comparative Evaluation Benchmarked Against Active Learning and an Independent Expert Reviewer

Robert D. Keeling • Ray Mangum • F. Eli Nelson • Kevin A. Reiss

Redgrave LLP {rkeeling, rmangum, enelson, kreiss}@redgravellp.com

June 2026

Abstract

We compare a generative-AI document-review workflow (Relativity `aiR for Review`) against a contemporaneous managed review using continuous active learning (Relativity Active Learning, or `RAL`), on a deliberately demanding responsiveness standard requiring application of multiple federal regulatory frameworks. Both workflows were run against the same 45,004-document corpus from the Mallinckrodt collection in the UCSF Opioid Industry Documents Archive. We then drew a 1,000-document simple random sample and obtained ground truth from a single subject-matter expert in a blinded review (73 responsive). `aiR for Review` achieved 88% recall (95% Wilson CI: 78.2–93.4) against `RAL`'s 64% (95% CI: 52.9–74.4); the intervals do not overlap. `aiR for Review`'s recall advantage came at the cost of 11 percentage points of precision (29% vs. 39%). Elusion rates—the fraction of truly responsive documents missed in the discard pile—were 1% for `aiR for Review` and 3% for `RAL`. A secondary informed re-review surfaced 10 additional responsive documents that the expert had missed in his blinded assessment but agreed were responsive after seeing `aiR for Review`'s rationale. The findings provide empirical support for the defensibility of a properly prompted generative-AI workflow under the reasonableness and proportionality standards established by case law and the Sedona Conference TAR 1 Reference Model.

Keywords: technology-assisted review, generative AI, e-discovery, recall, precision, TAR, defensibility

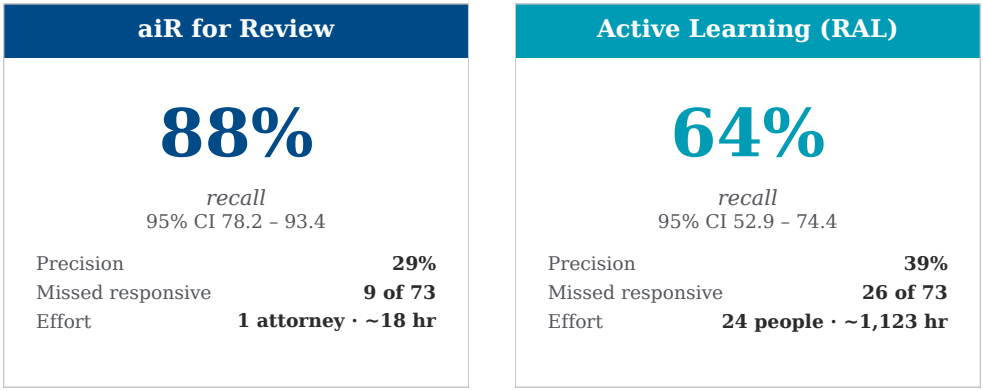

Figure 1. Headline comparison of the two workflows on a 1,000-document random sample (73 responsive by blinded expert ground truth). `aiR for Review` achieved 88% recall against `RAL`'s 64%—a 23-percentage-point advantage, with the 95% Wilson CIs not overlapping—at the cost of 11 percentage points of precision.

1 Introduction

The volume and complexity of electronically stored information (ESI) in civil litigation have grown dramatically over the past two decades. Organizations now generate vast quantities of digital records across email, messaging platforms, cloud-based collaboration tools, and enterprise applications, producing document collections in large matters that routinely reach into the millions [1]. The cost of reviewing these collections—traditionally performed by teams of dozens to hundreds of contract attorneys—remains among the largest expenses in modern litigation and government investigations. As ESI volumes continue to increase, the mismatch between the scale of document populations and the capacity of manual review has driven the legal profession's adoption of increasingly sophisticated technology solutions—originally transitioning from manual review to keyword searching to technology-assisted review (TAR).

For over a decade, defensible TAR workflows in litigation have relied on classifiers trained on labeled examples [1, 2]. TAR uses supervised learning algorithms to categorize documents into classes—here responsive to document requests—based on similar document examples [3]. More specifically, TAR is a process by which "computers are programmed to search large quantities of documents . . . to mimic the document selection process of a knowledgeable, human document review" [4].

The rapid emergence of large language models (LLMs) has introduced a fundamentally new approach. Rather than relying on labeled training examples to build statistical classifiers, generative AI offers an alternative: LLM-based review systems accept the natural-language review protocol similar to what would be provided to a team of human reviewers and apply it directly to classify documents.

The Sedona Conference's TAR 1 Reference Model [5] treats both TAR and GenAI review as instances of a common defensibility framework. Whether generative-AI review meets the recall and proportionality standards courts have applied to traditional TAR is, however, an empirical question. While early empirical work has demonstrated the viability of this approach, these studies have identified substantial avenues for improvement [6].

This paper fills in that gap—reporting a head-to-head comparison of Relativity's `aiR for Review` against an industry-standard TAR continuous active learning workflow using Relativity Active Learning (`RAL`) on a single matter under a single responsiveness standard chosen for its complexity: a document was responsive only if it contained evidence of compliance with—or violation of—federal requirements governing pharmaceutical marketing and the handling of controlled substances. Reference to the regulated subject matter was not, on its own, sufficient: the reviewer—human or model—had to apply the governing legal standards to the document's content. Most TAR benchmarks treat any document discussing a topic as responsive; this task is meaningfully harder.

We ran both workflows (`aiR for Review` and `RAL`) against the same 45,004-document corpus from the Mallinckrodt collection in the UCSF Opioid Industry Documents Archive [7], driven by the same review protocol. The `RAL` workflow was a first-pass managed review using the `RAL` classifier: a 24-person team—21 contract-attorney reviewers, a review manager, a team lead, and dedicated QC reviewers—reviewed the corpus over seven business days under a continuous active learning approach.

The generative-AI workflow used the same review protocol as the starting point for a prompt run through Relativity's `aiR for Review`; a single attorney iterated the prompt over approximately 18 hours and applied the finalized version to the corpus in a single processing run. Other than the common review protocol, neither workflow shared information with the other.

To establish ground truth for benchmarking performance, we drew a 1,000-document simple random sample from the corpus and asked Kevin A. Reiss, Counsel at Redgrave LLP and the study's subject-matter expert, to review every document blindly, without access to either workflow's predictions. His labels (73 Responsive, 927 Not Responsive) define the ground truth against which we measure each workflow's recall, precision, and elusion.

Each workflow's predictions were then measured against this ground truth. After the primary measurement, the expert returned to 151 documents where his blinded assessment had disagreed with `aiR for Review`'s prediction and re-examined each with `aiR for Review`'s rationale and citations visible. The re-review tested whether `aiR for Review` had identified responsive material that the blinded assessment missed.

We find:

- **Recall.** `aiR for Review` achieved 88% recall (95% Wilson CI [8]: 78.2–93.4), identifying 64 of 73 responsive documents; `RAL` achieved 64% (52.9–74.4), identifying 47. The intervals do not overlap. Elusion was 1% for `aiR for Review` and 3% for `RAL`, a difference of roughly $3\times$. The recall figures place a properly prompted generative-AI workflow within the range courts have accepted as reasonable for technology-assisted review under the proportionality standard [9].
- **Precision in context.** `aiR for Review` gained 23 percentage points of recall against a cost of 11 percentage points of precision (29% vs. 39%). Both precision figures are depressed by the corpus's low richness (7%, well below the 20–25% TAR design target), a mathematical property of any binary classifier on a sparse-positive corpus rather than a defect of either workflow (Section 5.2). Holding observed recall and false positive rate constant, projected precision rises to 56% for `aiR for Review` and 67% for `RAL` at 20% richness, the range typical of TAR validation studies.
- **Discovery of overlooked documents.** The expert overturned 10 of his blinded Not Responsive labels to Responsive after reviewing `aiR for Review`'s rationale—documents he himself ultimately agreed were responsive after seeing the model's reasoning. `aiR for Review` can surface responsive material that even an experienced reviewer initially overlooks.
- **Population-scale review burden.** Extrapolated to the full 45,004-document corpus, `aiR for Review` would flag roughly 9,971 documents against `RAL`'s 5,019—an additional 4,952 to review—while missing roughly 376 responsive documents against `RAL`'s 1,177, a net recovery of about 801 responsive documents. The `aiR for Review` workflow consumed approximately 18 hours from one attorney; the `RAL` workflow consumed approximately 1,123 hours from a 24-person team over seven business days.

Paper organization. Section 2 describes the corpus, review topic, workflows, and validation protocol. Section 3 specifies the prediction rules and ground-truth derivation. Section 4 presents raw classification outcomes and the agreement structure between the two workflows. Section 5 reports performance metrics with 95% Wilson confidence intervals and situates them in the legal-standards context. Section 6 details the informed re-review. Section 7 visualizes classification flows. Section 8 compares review effort. Appendix A extrapolates the sample-level performance to the full population.

Key metrics used in this report

Recall (completeness). Of all truly responsive documents, what fraction did the review process identify?

Precision (efficiency). Of the documents flagged as responsive, what fraction actually were?

Elusion (risk in the discard pile). Of documents classified as not responsive, what fraction were actually responsive?

Richness (prevalence). The fraction of documents in a collection that are responsive. Lower richness makes accurate classification harder for all review processes.

Confidence interval (uncertainty). A 95% confidence interval (CI) is a range that would contain the true population value in 95% of repeated samples.

2 Study Background

2.1 Document corpus

The study population comprises 45,004 documents drawn from the Mallinckrodt corpus in the UCSF Opioid Industry Documents Archive [7], a publicly disclosed collection. It contains a mix of emails, Word documents, Excel spreadsheets, PowerPoint presentations, and other file types. A review protocol was synthesized from one of the initial complaints in the litigation, simulating a new document request issued against the produced population. A separate control set of 250 randomly selected documents established an initial population richness estimate of 6% (16 of 250 responsive). That figure falls well below the 20–25% design target and reflects the smaller pool of documents responsive to a compliance/violation standard than to topics where any document "related to" the subject is responsive. Rather than substitute a simpler classification task to hit the richness target, the team chose to proceed with the experiment as designed, accepting the lower richness as a realistic feature of more demanding review topics sometimes encountered in practice. The same control set was also used to iteratively refine the `aiR for Review` prompt.

2.2 Review topic

For purposes of this experiment, responsiveness was defined as follows:

> A document is responsive if it contains evidence of Mallinckrodt's or Covidien's compliance with, violations of, or reckless disregard of federal laws and regulations regarding how pharmaceutical companies may market or promote their products, including any applicable statutory or regulatory requirements for controlled substances as appropriate. This includes affirmative steps taken by Mallinckrodt or Covidien to ensure compliance with applicable federal laws and

> regulations, including documentation or observations of non-compliant actions or status, or notable omissions of compliance requirements where inclusion of these details would be reasonably expected (e.g., as part of training programs for sales representatives). A document is not responsive merely because it discusses or relates to sales activities or drug promotion generally.

The review protocol provided to human reviewers and the `aiR for Review` prompt both included substantial additional context beyond the definition above—including background on the applicable regulatory regimes and how to apply them—to help reviewers and the model evaluate documents against the governing legal standards.

2.3 Review workflows

This analysis is limited to the two technology-assisted workflows described below: `aiR for Review` and `RAL`. We detected no population-level trends that would undermine the conclusions reported here.[1]

`aiR for Review` **(GenAI TAR 1 workflow).** Ray Mangum, Partner at Redgrave LLP, developed the `aiR for Review` prompt. After 43 prompt iterations—many brief and exploratory—spanning approximately 18 hours across multiple sessions, the finalized prompt was applied to the full 45,004-document population (January 2026). This workflow follows the GenAI TAR 1 process described in the Sedona Conference TAR 1 Reference Model: an attorney writes and iteratively refines a natural-language prompt instructing the LLM to classify each document, with performance measured against a random control set.

`RAL` **(Relativity Active Learning—TAR 2/CAL workflow).** A first-pass managed review staffed by 21 contract-attorney reviewers, a review manager, a team lead, and dedicated QC reviewers, run through Relativity Active Learning in Review Center—Relativity's review-management tool that builds custom review queues, uses AI to prioritize relevant documents, and provides a reporting dashboard for tracking review state and productivity. The team was trained and calibrated against the review protocol, then reviewed the population over seven business days under a continuous active learning (CAL) workflow. Cimplifi, the managed-review vendor, achieved 11% richness in the prioritized queue. Remaining documents falling below the `RAL` relevance cutoff were manually reviewed; post-cutoff responsiveness review quickly reached 0%. In this report, `RAL` refers to the combined output of this full, first-pass workflow—inseparable from the review team's performance, the vendor's QC design, and the calibration process—and the measured recall reflects this specific implementation rather than the capability of continuous active learning in general.

2.4 Validation

Kevin A. Reiss, Counsel at Redgrave LLP, served as the sole subject-matter expert (SME) for all ground-truth labeling, contract-reviewer calibration, and validation review. Mr. Reiss has experience across all aspects of discovery in government in-

[1] The full corpus was also subjected to manual review as part of the broader experiment; analysis of that workflow is reserved for future work.

vestigations and state and federal civil and criminal litigation, including managing teams of over 100 document reviewers.

A simple random sample of 1,000 documents—separate from the 250-document control set used during prompt development—was drawn from the full population to serve as the independent validation sample. Mr. Reiss adjudicated each document in a blinded review, without seeing `aiR for Review`'s scores or rationale, and his blinded assessments define the ground truth for this study. A subsequent informed re-review (Section 6) provided a robustness check on that ground truth and demonstrates `aiR for Review`'s ability to surface responsive documents that the expert initially overlooked.

Where methodological choices could favor one workflow over the other, this study consistently resolved them against `aiR for Review` to conservatively report performance accuracy. For example, error documents were counted as `aiR for Review` positive predictions, ground truth was defined solely by the expert's blinded assessment, and the informed re-review results are reported only as a secondary analysis and not used to adjust the initial ground-truth determinations.

3 Study Design and Ground Truth

3.1 Prediction rules

Each review process classified documents as Predicted Responsive or Predicted Not Responsive according to the following rules.

`aiR for Review`. `aiR for Review` assigned each document a score reflecting its assessment of responsiveness. Documents scored *Very Responsive* (4), *Responsive* (3), or *Borderline* (2) were treated as Responsive predictions—any document `aiR for Review` judged to have at least a plausible basis for responsiveness was counted in the Responsive set for purposes of this experiment. *Error* documents (score -1) were system failures that could not be scored; they were automatically routed to human review and are likewise counted as Responsive predictions because they entered the review queue, and are not treated as misses for purposes of measuring recall.

Note on error documents

The 12 error documents are counted as `aiR for Review` Responsive predictions because they entered the review queue. Of these 12, 9 have ground truth = Not Responsive and contribute to `aiR for Review`'s false positive count; the remaining 3 have ground truth = Responsive and are true positives. This treatment is consistent with the decision to resolve methodological choices against `aiR for Review`.

`RAL`. A document was predicted Responsive when the first-pass, managed-review workflow's final label was Responsive. This includes documents prioritized by `RAL` and 240 documents that fell below the relevance cutoff but were nonetheless manually reviewed; of those 240, 2 were ultimately labeled Responsive.

3.2 Ground-truth derivation

Ground truth was defined by the expert's independent blinded assessment of all 1,000 documents in the sample. The blinded review was conducted without access to `aiR for Review`'s scores or rationale; those blinded assessments were the sole basis for ground truth.

Ground truth by construction

Because the expert's blinded assessment defines ground truth, the expert has perfect recall and precision by construction; his labels are the standard against which `aiR for Review` and `RAL` are measured. The informed re-review (Section 6) provides an additional check on the stability of these judgments, ultimately revealing 10 documents that the expert himself later agreed were Responsive after reviewing `aiR for Review`'s rationale and citations. These 10 documents were not folded back into the ground truth used for the primary metrics; recall, precision, and elusion throughout this paper are computed against the SME's original blinded labels. The informed re-review was conducted after the primary measurement as a secondary analysis and as groundwork for possible future work.

3.3 Sample composition

Table 1. Validation sample composition (blinded ground truth).

Blinded assessment	Count
Responsive	73
Not Responsive	927
Total	1,000

Sample design

The 1,000-document sample was drawn randomly from the full population and includes natural proportions of agreements, conflicts, borderlines, and errors. The sample size provides approximately $\pm 10\%$ confidence intervals for recall estimates and is sufficient to demonstrate with 95% confidence that `aiR for Review` achieved at least 80% recall.

Table 2. `aiR for Review` score distribution on the 1,000-document sample.

Score	Description	Count	`aiR for Review` Prediction
4	Very Responsive	1	Responsive
3	Responsive	61	Responsive
2	Borderline	150	Responsive
1	Not Responsive	682	Not Responsive
0	Junk	94	Not Responsive
-1	Error	12	Responsive (sent to review)

4 Sample Results

4.1 Sample-level overview

With the study design established, we turn to the raw classification outcomes for each review process. Tables 3 and 4 report raw classification counts on the 1,000-document sample.

Table 3. Sample-level prediction totals.

Review process	Predicted Responsive	Predicted Not Responsive	Total
`aiR for Review`	224	776	1,000
`RAL`	120	880	1,000

Table 4. Confusion matrices for `aiR for Review` and `RAL` against blinded ground truth.

Ground truth	Outcome	`aiR for Review`	`RAL`	Subtotal
Responsive (73)	Correctly identified (TP)	64	47	
	Missed (FN)	9	26	
Not Responsive (927)	Correctly identified (TN)	767	854	
	Over-identified (FP)	160	73	
Total		1,000	1,000	

Abbreviations. TP = True Positive (predicted Responsive, actually Responsive); FN = False Negative (predicted Not Responsive, actually Responsive); TN = True Negative (predicted Not Responsive, actually Not Responsive); FP = False Positive (predicted Responsive, actually Not Responsive).

4.2 Classification detail and agreement

The table below presents classification outcomes from two perspectives. The first section shows how `aiR for Review` and `RAL` agreed or disagreed on each document and how the expert's ground truth resolved each category. The second section rolls up each approach's confusion matrix into prediction-level summaries. The *% Responsive* column shows the fraction of documents in each row that were actually responsive.

Distribution of ground truth responsive documents. Of the 73 documents the expert labeled Responsive in his blinded assessment, 44 were identified by both `aiR for Review` and `RAL`, 20 were found only by `aiR for Review` (`RAL` missed them), and 3 were found only by `RAL` (`aiR for Review` missed them). Six responsive documents were missed by both processes.

Summary of disagreement patterns. The agreement analysis reveals a consistent asymmetry: when `aiR for Review` and `RAL` disagreed, `aiR for Review` overwhelmingly erred toward over-inclusion—flagging documents as responsive that the expert deemed not responsive. `RAL`'s disagreements ran in the opposite direction—it tended to discard documents that turned out to be responsive.

Table 5. Agreement structure between `aiR for Review` and `RAL` predictions, resolved by ground truth.

Category	Count	Actually Responsive	Actually Not Responsive	% Responsive
Agreement analysis				
Both Responsive	79	44	35	56%
Both Not Responsive	735	6	729	0.8%
`aiR for Review` only	145	20	125	14%
`RAL` only	41	3	38	7%
`aiR for Review` vs. ground truth				
Predicted Responsive	224	64	160	29%
Predicted Not Responsive	776	9	767	1.2%
`RAL` vs. ground truth				
Predicted Responsive	120	47	73	39%
Predicted Not Responsive	880	26	854	3%
Total	1,000	73	927	7%

`aiR for Review` false positive decomposition (160 total)

All 160 false positives are documents that `aiR for Review` predicted Responsive but the expert's blinded assessment classified as Not Responsive. These include 9 error documents (score $= -1$, routed to review). Section 6 examines the 151 non-error false positives through the informed re-review, which finds that 10 of them were later confirmed Responsive by the expert himself.

5 Performance Analysis

The raw counts in Section 4 establish what each review process got right and wrong. This section translates those counts into standard performance metrics and quantifies the uncertainty around each estimate.

5.1 Sample performance metrics

Abbreviations. PPV (Positive Predictive Value) = Precision; Sensitivity = Recall; FPR (False Positive Rate) = fraction of actually Not Responsive documents incorrectly predicted Responsive; FNR (False Negative Rate) = fraction of actually Responsive documents incorrectly predicted Not Responsive.

`RAL` recall: workflow QC vs. population recall

The first-pass `RAL` workflow's TAR tool reported 100% recall based on an elusion sample drawn from its own discard pile. We measured 64% recall through a blinded expert review of a random sample drawn from the full population. The figures are not contradictory; they answer different questions.

The TAR tool's number comes from sampling the algorithm's discard pile and checking whether any of the discarded documents were actually responsive. None were, so the tool reported 100% recall—an accurate statement about the discards. Our 64% comes

Table 6. Sample-level performance metrics.

Metric	Formula	Substitution	Percent
aiR for Review			
Precision (PPV)	TP / (TP + FP)	64 / 224	29%
Recall (Sensitivity)	TP / (TP + FN)	64 / 73	88%
Accuracy (point est.)	(TP + TN) / n	831 / 1,000	83%
Elusion (point est.)	FN / (FN + TN)	9 / 776	1%
FPR	FP / (FP + TN)	160 / 927	17%
FNR	FN / (FN + TP)	9 / 73	12%
RAL			
Precision (PPV)	TP / (TP + FP)	47 / 120	39%
Recall (Sensitivity)	TP / (TP + FN)	47 / 73	64%
Accuracy (point est.)	(TP + TN) / n	901 / 1,000	90%
Elusion (point est.)	FN / (FN + TN)	26 / 880	3%
FPR	FP / (FP + TN)	73 / 927	8%
FNR	FN / (FN + TP)	26 / 73	36%

from a different vantage point: a blinded review by the subject-matter expert of a random sample of every document in the collection, regardless of where the workflow had routed it. It asks how many of all responsive documents the workflow surfaced. The two diverge because the discard pile is not the only place a responsive document can be lost. In an active-learning workflow, the algorithm presents documents to contract reviewers, who code each as Responsive or Not Responsive; the algorithm then trains on those codings. When a reviewer wrongly codes a truly responsive document as Not Responsive, the document drops out of the production set, and the algorithm treats the label as ground truth for its next round of training. Elusion sampling cannot detect those reviewer-caused misses because it samples only from the algorithm's discards. The 64% population recall therefore reflects the combined performance of the contract-reviewer team, vendor QC, and calibration process—not a property of the algorithm in isolation.

5.2 Precision in context: the effect of richness

Precision depends not only on the quality of the classifier but also on the richness (prevalence) of responsive documents in the collection. When richness is low, even a classifier with a very low false positive rate will flag many more non-responsive documents than responsive ones—simply because non-responsive documents vastly outnumber responsive ones. This is a mathematical property of all binary classifiers.

Two metrics that are intrinsic to a classifier—independent of richness—are recall and the false positive rate. Given those two rates, the precision that would be observed at any richness level can be computed with a standard formula:

$$\text{Precision} = \frac{\text{Recall} \times \text{Richness}}{\text{Recall} \times \text{Richness} + \text{FPR} \times (1 - \text{Richness})} \tag{5.1}$$

Table 7. Projected precision at varying richness.

Richness	`aiR for Review` Precision	`RAL` Precision
7% (observed)	29%	39%
10%	36%	48%
20%	56%	67%
30%	69%	78%
40%	77%	85%

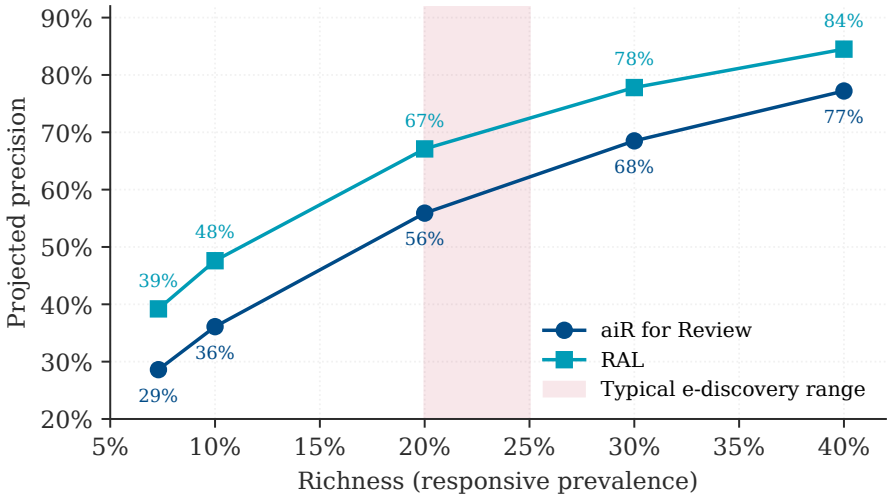

Figure 2. Projected precision at varying richness, holding each workflow's observed recall and false positive rate constant. The shaded band marks the 20–25% richness range typical of many e-discovery reviews.

Topic complexity

Population richness was 7% (validation sample: 73 of 1,000 responsive; the `RAL` control set independently estimated 6%, 16 of 250), well below the 20–25% design target. The review topic spans application of multiple regulatory frameworks—pharmaceutical marketing, controlled-substance requirements, and compliance programs—making consistent human coding inherently difficult and depressing absolute precision for all review processes. This context matters when interpreting the precision figures above.

5.3 Legal context for performance standards

The recall and precision figures reported here should be evaluated against the applicable legal standards. Courts have consistently held that producing parties need not achieve perfection—the standard is whether the review process was reasonable and proportional to the needs of the case. *See, e.g.,* [10]; [11]; [12].

The Sedona Conference TAR 1 Reference Model establishes that generative AI review follows the same defensible process steps as traditional discriminative TAR: *Scope*, *Label Control Set*, *Iterate Model*, *Classify*, and *Validate*. This study provides empirical validation as contemplated by that framework. On this corpus and review topic,

the `aiR for Review` workflow produced materially fewer false negatives than this managed-review implementation, under a random-sample validation design—with recall figures meeting or exceeding the range that courts have accepted as reasonable for technology-assisted review [1, 2].

5.4 Precision and recall with 95% confidence intervals

Because the 1,000-document sample is drawn from a larger population, these point estimates carry sampling uncertainty. A confidence interval quantifies that uncertainty: a 95% CI means that, under repeated sampling, 95% of such intervals would contain the true population proportion.

We use the Wilson score interval [8], which is well-suited for binomial proportions—particularly when the true proportion is near 0 or 1, or when sample sizes are moderate.

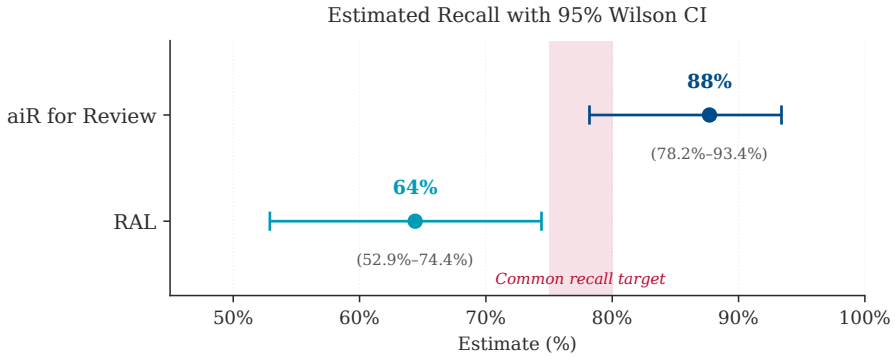

Figure 3. Estimated recall with 95% Wilson confidence intervals. The shaded band marks the 75–80% range commonly negotiated as a recall target in ESI protocols. The 95% CIs do not overlap, confirming a statistically significant difference.

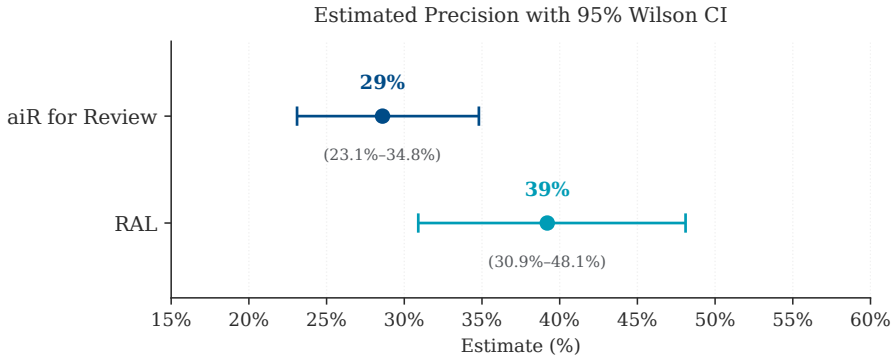

Figure 4. Estimated precision with 95% Wilson confidence intervals.

Table 8. Wilson 95% confidence interval components for `aiR for Review` and `RAL`.

Process	Metric	p	Denom.	Center	Margin	Lower	Upper
`aiR for Review`	Precision	0.286	1.020	0.289	0.059	23.1%	34.8%
`aiR for Review`	Recall	0.877	1.053	0.858	0.076	78.2%	93.4%
`RAL`	Precision	0.392	1.032	0.395	0.086	30.9%	48.1%
`RAL`	Recall	0.644	1.053	0.637	0.107	52.9%	74.4%

6 Informed Re-Review

The primary analysis (Sections 2 to 5) measures `aiR for Review` and `RAL` against the expert's independent blinded assessment. This secondary analysis examines what happened when the expert was subsequently shown `aiR for Review`'s rationale for the documents he had initially classified as Not Responsive.

6.1 Re-review methodology

After the blinded review was complete, 151 documents where the expert had coded Not Responsive but `aiR for Review` had predicted Responsive (scores 2, 3, or 4) were flagged for informed re-review. The expert re-examined each document with `aiR for Review`'s rationale, considerations, and document citations visible, and recorded one of three outcomes:

Table 9. Informed re-review outcomes (151 documents).

Re-review outcome	Count	Expert's informed assessment
Agree (expert changed to Responsive)	10	Expert confirmed Responsive
Borderline	6	Close to threshold; ultimately Not Responsive
Disagree (expert maintained NR)	135	Expert maintained Not Responsive
Total re-reviewed	151	

Directionality

Because the informed re-review was one-directional—only documents that the expert had coded Not Responsive and `aiR for Review` had predicted Responsive were re-examined—it captures `aiR for Review`'s ability to find false negatives in the expert's blinded assessment but does not test whether `aiR for Review`'s Not Responsive predictions might also be incorrect. The primary analysis in Sections 2 to 5, which measures recall and precision against the blinded ground truth, provides the complete performance picture.

6.2 Results

Of the 151 documents re-reviewed, the expert overturned 10 of his original Not Responsive calls to Responsive—an overturn rate of 7%. These are documents that the expert himself confirmed were Responsive after seeing `aiR for Review`'s analysis.

6.3 Implications

The fact that these 10 were overturned because of supplied rationale demonstrates that `aiR for Review` can surface responsive material with analytical rigor that even an experienced reviewer can find persuasive. In production settings, this capability supports multiple deployment models: `aiR for Review` as the primary review classifier with human confirmation of flagged documents, as a second-look layer that identifies documents a first-pass review may have missed, or as a direct augmentation of expert review, with the attorney reviewing `aiR for Review`'s flagged predictions.

7 Document Flow Visualizations

The flow diagrams below trace how each process's predictions move from the 1,000-document sample through to ground-truth outcomes. The width of each band is proportional to the number of documents following that path.

7.1 `RAL` classification (first-pass review)

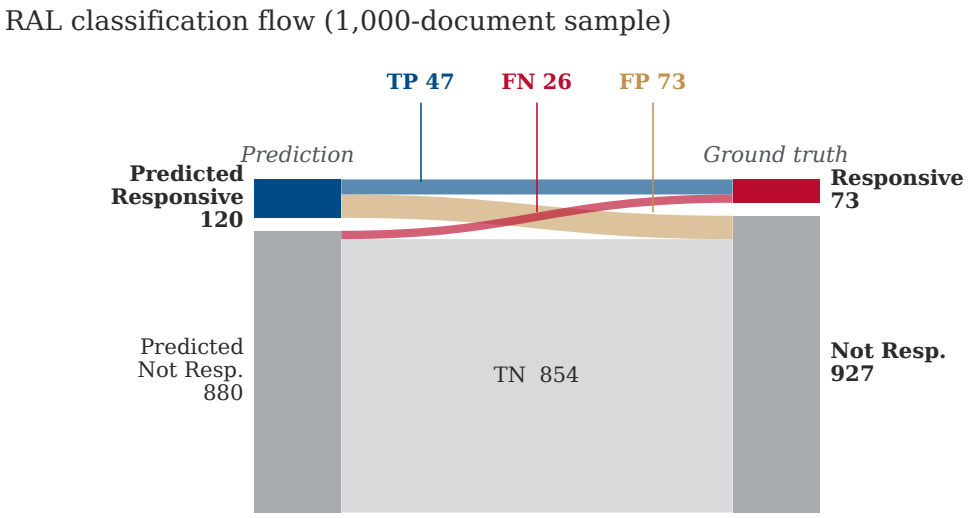

Figure 5. `RAL` classification flow on the 1,000-document validation sample. The narrow red band represents the 26 false negatives.

7.2 `aiR for Review` classification

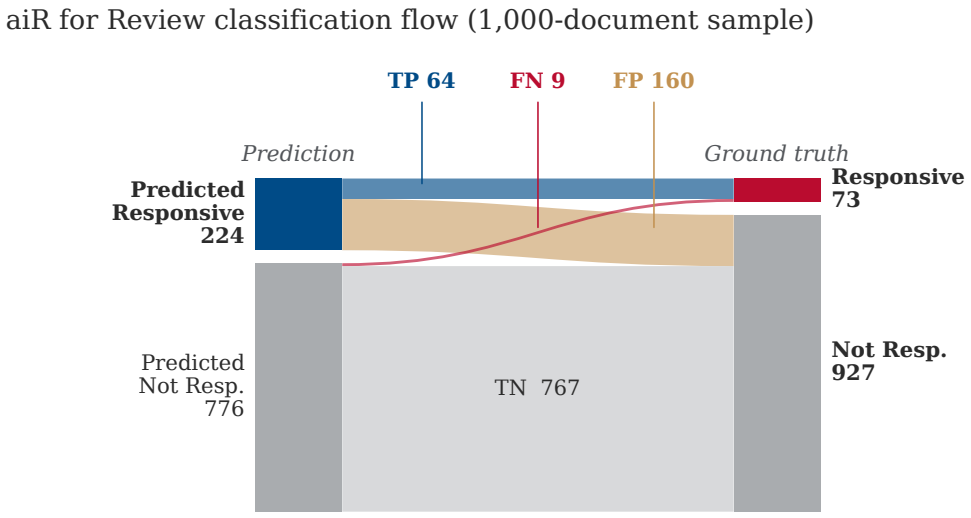

Figure 6. `aiR for Review` classification flow. Compared with `RAL`, the false negative band is narrower (9 vs. 26) while the false positive band is wider (160 vs. 73).

In this study, the two diagrams illustrate the fundamental trade-off: `aiR for Review`'s broader net caught more responsive documents at the cost of additional review, while `RAL`'s narrower classification missed a larger share of responsive material.

8 Review Effort Comparison

8.1 Effort summary

The `aiR for Review` workflow required approximately 18 hours from one attorney; the `RAL` workflow required approximately 1,123 hours from a 24-person team over seven business days. Table 10 reports only the attorney labor in each workflow. Once the finalized `aiR for Review` prompt was applied to the corpus, the model processed all 45,004 documents in a single run lasting approximately one hour of elapsed time; that processing time is not attorney labor and is not included in the table.

Table 10. Effort comparison between the two workflows.

Category	`aiR for Review` **Workflow**	`RAL` **Workflow**
Prompt Drafting and Iteration	∼18 hours	—
Review Manager	—	45.5 hours
Team Lead	—	35.75 hours
Contract Reviewers (21)	—	973.75 hours
Contract Reviewer QC	—	68 hours
Total*	∼18 hours	∼1,123 hours
Personnel*	1 attorney	24 people

*Totals and personnel counts reflect only the attorney labor performed within each workflow. Kevin A. Reiss, the study's subject-matter expert, drafted the common review protocol that drove both workflows (approximately 3 hours of labor); that labor is not included in either workflow's total or personnel count. The `RAL` hours reflect the active learning review phase only and are reported by Cimplifi, the managed-review vendor. The 973.75 contract-reviewer hours include an estimated 42 hours of onboarding and training (approximately 2 hours on the first day for each of 21 reviewers), with an additional 2 hours each for the Review Manager and Team Lead.

8.2 Review coverage

During the active learning phase, contract reviewers worked through 34,409 of the 45,004 documents in the study population (77%), prioritized by the `RAL` relevance model in descending order of likely responsiveness. The remaining 10,595 documents fell below the relevance cutoff and were reviewed separately over a two-day period. Of those 10,595 below-cutoff documents, only 12 were ultimately labeled Responsive. By contrast, `aiR for Review` classified the entire 45,004-document population in a single processing run.

A Estimated Population Performance

Because the 1,000-document sample was drawn at random, the sample performance rates are unbiased point estimates of the full-population performance. Applying `aiR for Review`'s observed prediction rate of 22% to the 45,004-document population yields an estimate of approximately 9,971 documents flagged. Applying `RAL`'s observed prediction rate of 12% yields approximately 5,019. Applying each workflow's false negative rate to the estimated 3,305 responsive documents in the population yields approximately 376 missed by `aiR for Review` versus approximately 1,177 missed by `RAL`—a difference of roughly 801 additional responsive documents recovered by `aiR for Review`.

References

[1] Maura R. Grossman and Gordon V. Cormack. Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law & Technology*, 17:11, 2011. Available at https://scholarship.richmond.edu/jolt/vol17/iss3/5.

[2] Gordon V. Cormack and Maura R. Grossman. Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*, pages 153–162, New York, NY, 2014. ACM. doi: 10.1145/2600428.2609601. Available at https://dl.acm.org/doi/pdf/10.1145/2600428.2609601.

[3] Robert Keeling, Rishi Chhatwal, Nathaniel Huber-Fliflet, Jianping Zhang, and Haozhen Zhao. Using Machine Learning on Legal Matters: Paying Attention to the Data Behind the Curtain. *Hastings Science & Technology Law Journal*, 11:9, 2020.

[4] Charles Yablon and Nick Landsman-Roos. Predictive Coding: Emerging Questions and Concerns. *South Carolina Law Review*, 64:633, 2013.

[5] Tara Emory, Jeremy Pickens, and Wilzette Louis. TAR 1 Reference Model: An Established Framework Unifying Traditional and GenAI Approaches to Technology-Assisted Review. *Sedona Conference Journal*, 25:109, 2024. Available at https://www.thesedonaconference.org/download-publication?fid=7301.

[6] Eugene Yang et al. Beyond the Bar: Generative AI as a Transformative Component in Legal Document Review. In *IEEE International Conference on Big Data*, December 2024. Available at https://ieeexplore.ieee.org/document/10826089/.

[7] UCSF Industry Documents Library. Opioid Industry Documents Archive: Mallinckrodt Collection. https://www.industrydocuments.ucsf.edu/opioids/, 2026. Accessed 2026.

[8] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. Available at https://www.jstor.org/stable/2276774.

[9] Maura R. Grossman and Gordon V. Cormack. Comments on "The Implications of Rule 26(g) on the Use of Technology-Assisted Review". *Federal Courts Law Review*, 7:285, 2014. Available at https://www.fclr.org/fclr/articles/pdf/comments-implications-rule26g-tar-62314.pdf.

[10] *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125 (S.D.N.Y. Mar. 2, 2015) (Peck, M.J.). Available at https://www.thesedonaconference.org/sites/default/files/judicial_opinions/Rio%20Tinto%20PLC%20v.%20Vale%20S.A.%2C%20306%20F.R.D.%20125.pdf.

[11] *Hyles v. City of New York*, No. 10 Civ. 3119-AT-AJP, 2016 WL 4077114 (S.D.N.Y. Aug. 1, 2016) (Peck, M.J.). Available at https://www.thesedonaconference.org/sites/default/files/judicial_opinions/Hyles%20v%20New%20York%20City%2008-01-16%20Pacer.pdf.

[12] Fed. R. Civ. P. 26(b)(1) (2015 amendments) and accompanying Committee Notes.

Suggested citation. Robert D. Keeling, Ray Mangum, F. Eli Nelson, & Kevin A. Reiss, *Generative AI for Complex Document Review: A Comparative Evaluation Benchmarked Against Active Learning and an Independent Expert Reviewer*, Redgrave LLP Working Paper 2026-01 (June 2026).