

# Comparing GenAI to Machine Learning and Manual Review for eDiscovery

Gordon V. Cormack<sup>1</sup>, Maura R. Grossman<sup>1</sup>, and Jeremy Pickens<sup>2</sup>

<sup>1</sup> University of Waterloo, ON, Canada

<sup>2</sup> Elevate, USA

**Abstract.** Keeling et al. [9] report on a case study recently conducted by Redgrave LLP in which they assert that Relativity’s Generative AI Review tool (aiR) achieved 88% recall while Relativity Assisted Learning (RAL), its machine-learning tool, achieved 64% recall. These results are not comparable because the methods employed have different objectives, because of inter-assessor disagreement, and because they represent different recall-precision and cost-benefit trade-offs. Our analysis of the Redgrave data indicates that, when these differences are taken into account, aiR and manual review are comparable. Furthermore, established machine-learning models employing Continuous Active Learning (CAL) show state-of-the-art effectiveness, and superior efficiency, when trained only on a sample of 2,000 first-pass-reviewer responsiveness determinations.

**Keywords:** Technology-assisted review · e-discovery · relevance assessment · information retrieval evaluation · continuous active learning

## 1 Introduction

Keeling et al. [9] conducted a head-to-head case study for Redgrave LLP (the Redgrave Case Study) comparing Relativity aiR Review (aiR), a generative AI tool, with an established managed review process employing Relativity Assisted Learning (RAL) [9]. Keeling et al. reported that Relativity’s aiR achieved 87.7% recall and 28.6% precision, compared to RAL’s 64.4% recall and 39.2% precision.

These results – reported in multiple headlines – are incomparable for three principal reasons:

1. The RAL process identified for production those documents that were identified by RAL *and also coded responsive* by a team of junior “contract lawyers” (the L1 team). In contrast, the aiR process identified for production only those documents that were identified by aiR, *without such human review*. We call these methods (L1-) “filtered” and “unfiltered” TAR methods, respectively.
2. The L1 review team, the aiR prompt engineer, and the subject matter expert (SME), whose assessments were used for evaluation, have different conceptions of responsiveness, which will naturally influence recall and precision

estimates; results according to the SME may simply reflect how similar the L1 and prompt engineer teams’ conceptions of responsiveness were to the SME’s, rather than a technological difference.

3. The aiR and RAL results represent different recall-precision trade-offs. Choosing the best trade-off involves balancing proportionality factors such as completeness, cost, burden, risk, and timeliness. We cannot automatically say that one production is “better” than another without considering the risk and burden to both the producing and requesting parties, respectively.

This work shows that when the L1 review team has the final say on which documents are responsive (filtered TAR), or when all documents identified by the system are deemed responsive regardless of the L1 coding (unfiltered TAR), the differences in review effectiveness between aiR and managed review using RAL essentially vanish. Furthermore, decade-old supervised machine-learning protocols – e.g., Continuous Active Learning (CAL) [2] for filtered TAR, and Scalable Continuous Active Learning (S-CAL) [3] for unfiltered TAR – yield equal or better efficiency and effectiveness.

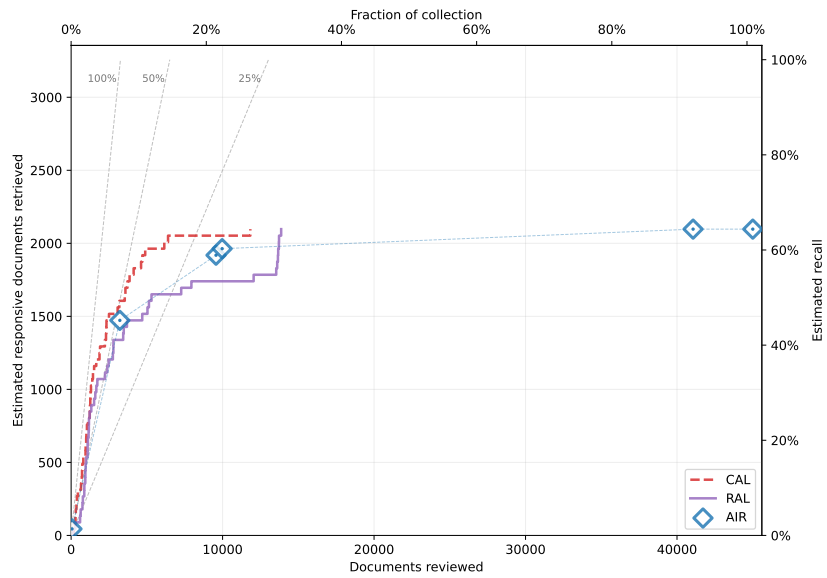
## 2 Dataset

The Redgrave Case Study used 45,004 documents from the UCSF Opioid Industry Documents Archive [1] (the Redgrave Dataset). RAL managed review involved manual coding for responsiveness of 34,409 documents in the order identified by RAL, followed later by manual coding of an additional 10,595 documents not identified by RAL, resulting in L1 coding for the entire Redgrave Dataset. The aiR review involved composing GenAI prompts, and having the GenAI label each document. Independent review by an SME of a simple random sample ( $n=1,000$ ) identified 73 documents as responsive. Treating the SME coding as ground truth, the RAL review achieved 64.4% recall (95% CI: 52.3%–75.3%) and 39.2% precision (30.4%–48.4%), while aiR achieved 87.7% recall (95% CI: 77.9%–94.2%) and 28.6% precision (22.8%–35.0%).

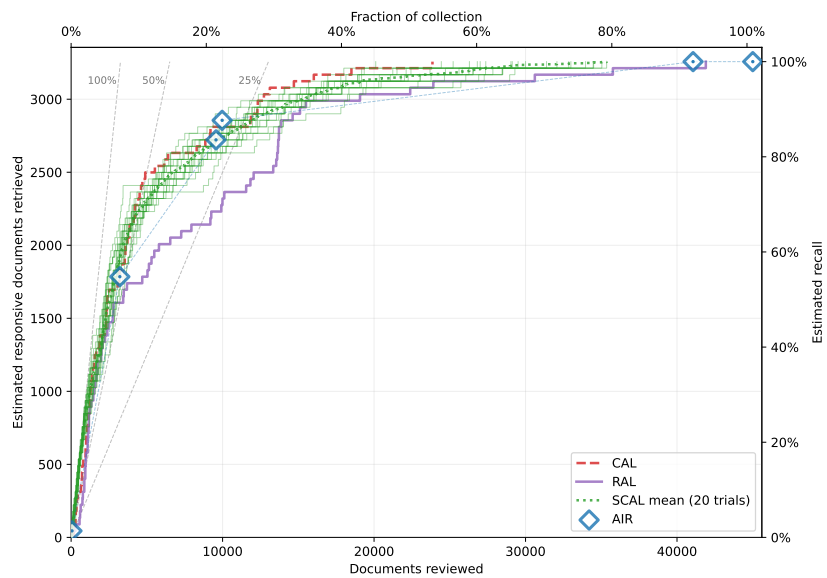
## 3 Filtered and Unfiltered TAR Comparisons

Also treating the SME coding as ground truth, we plotted gain curves for the filtered (Figure 1) aiR, RAL, and CAL methods, and for the unfiltered (Figure 2) aiR, RAL, CAL, and S-CAL methods. The CAL and S-CAL implementations were derived from the open-source Baseline Model Implementation (BMI)<sup>3</sup> used for the TREC 2015 and 2016 Total Recall Tracks [10, 8]. The CAL version was modified to retrain after each L1 coding decision, rather than in progressively larger batches (per Ghelani et al. [6]), and the S-CAL version was modified to require L1 coding for only a weighted statistical sample of 2,000 documents from the Redgrave Dataset, excluding the SME-coded sample of 1,000 documents.

<sup>3</sup> [plg.uwaterloo.ca/~gvcormac/trecvm/](http://plg.uwaterloo.ca/~gvcormac/trecvm/); implementation extracted from [plg.uwaterloo.ca/~gvcormac/trecvm/vmscripts.tar](http://plg.uwaterloo.ca/~gvcormac/trecvm/vmscripts.tar).



**Fig. 1.** Responsive documents identified by aiR, RAL, and CAL, filtered by L1 responsiveness coding.



**Fig. 2.** Responsive documents identified by aiR, RAL, CAL, and S-CAL, unfiltered by L1 responsiveness coding. Thin S-CAL curves show 20 individual trials; the thick curve shows their mean.

## 4 Assessor Agreement and Disagreement

Recall and precision are measured against a particular assessor’s judgments, not an infallible ground truth. In this case, the L1 reviewers coded as non-responsive 34% of the documents the SME deemed responsive, while the SME coded as non-responsive 61% of those that the L1 reviewers deemed responsive.

The agreement and disagreement between assessors is typically quantified using “overlap,” also known as Jaccard index: Overlap is the number of documents that a pair of assessors both label responsive, divided by the number that one or both label responsive.

**Table 1.** Inter-assessor overlap (Jaccard index) in the Redgrave Case Study.

|                     |       |
|---------------------|-------|
| SME & L1            | 0.322 |
| SME & S-CAL         | 0.361 |
| SME & aiR (liberal) | 0.275 |
| SME & aiR (strict)  | 0.421 |

**Table 2.** Intra- and Inter-assessor overlap (Jaccard index) reported in the literature.

| Retired Analysts [13] | DOJ Second Request [11] | State Archivist [4]                    |
|-----------------------|-------------------------|----------------------------------------|
| Primary & A 0.295     | Original & Team A 0.163 | Roger I & II (legal hold) 0.806        |
| Primary & B 0.350     | Original & Team B 0.158 | Roger I & II (public record) 0.602     |
|                       |                         | Roger I & II (restricted record) 0.642 |

Table 1 shows the overlap between SME coding and L1, S-CAL, and two aiR interpretations: *liberal* (including “marginal” and “error” categories, as necessary to achieve nearly 88% recall) and *strict* (including only “responsive” and “highly responsive” categories). The liberal and strict interpretations correspond to the second and fourth diamonds, respectively, in Figures 1 and 2. S-CAL overlap is calculated at its self-estimated recall-precision crossover.

Table 2 shows that reviewer disagreement is not unusual, ranging from 0.158 to 0.350 between professional reviewers, and from 0.602 to 0.806 between judgments made by the same SME at different times. Disagreement does not imply that either review is unreasonable. Managed L1 review is widely accepted as a reasonable basis for production decisions in the legal industry. Independent assessments, whether by L1 reviewers or SMEs, can support comparisons of retrieval methods performing the same task [12]. They do not establish an infallible ground truth.

The differences between results for the filtered and unfiltered tasks – both gain curves and recall-precision estimates – are largely explainable by inter-assessor disagreement. Within each task, the continuous gain curves come within

the sampling margin of error of the discrete aiR results, and the remaining differences – whether due to chance, training, or method – have little practical significance.

## 5 Proportionality and the Recall-Precision Trade-off

The doctrine of proportionality occasions us to balance the burden of reviewing or producing more documents, with their value. We can inform this analysis by examining the slope of the gain curve, if it is known. For RAL and CAL, the gain curve is constructed incrementally, and the utility of proceeding can be assessed by the frequency with which responsive documents are identified (i.e., the estimated slope of the gain curve). From S-CAL’s self-estimate of the entire gain curve, we may identify the point of diminishing returns. For aiR, we have only two viable cut points – liberal or strict – and no principled way to choose between them.

## 6 Limitations

The Redgrave Case Study considers a single matter, a single L1 review, and a single prompt engineering effort.

The Redgrave Dataset is a small, non-random subset of the 1.4M document Mallinckrodt collection, with an estimated richness of 7.3%, which may not be representative of the richness or nature of the documents in the larger collection (or other collections).

The opioids litigation has garnered a large amount of press and commentary, which may limit the applicability of results derived from it – or any similarly notorious dataset – to novel matters that are not equally publicized.

Redgrave’s aiR prompts and the methodology to develop them remain confidential.

Opaque aiR pricing precludes evaluation of cost in the proportionality analysis. Relativity documentation states that a subscription includes an allowance of AI “units” and that overage charges apply. When asked, Relativity declined to quantify the allowance or the overage charges.

## 7 Discussion and Conclusions

aiR can outperform RAL at specific points on the gain curve. But at the actual cutoff values – when comparing the methods using a common task – aiR does *not* outperform RAL. In the filtered TAR task, RAL achieves 64% recall (the highest possible given SME-L1 disagreement), while aiR achieves 59.5%. *CAL is superior to both at all points in the gain curve.* In the unfiltered TAR task, aiR achieves 88% recall, while RAL achieves 97.2%.

Using RAL (or CAL) for unfiltered TAR review would most likely be considered disproportionate because of review cost. S-CAL achieves a similar result

with a fixed review budget – in this case, 2,000 documents. S-CAL dominates RAL, and substantially matches both aiR-liberal and aiR-strict points. S-CAL, like RAL, exceeds 90% recall and eventually achieves 100% recall, which aiR never reaches.

At its self-calibrated recall-precision crossover, S-CAL agrees more closely with the SME than its L1 teacher does, consistent with prior results [5]. Self-calibration may be used to select other points on the gain curve, as occasioned by proportionality factors [3].

To assess scalability from the Redgrave Dataset to the full 1.4M-document Mallinckrodt collection, we consider cost and effectiveness. CAL and RAL review costs depend primarily on the number of responsive documents, not the collection size. Adding mostly non-responsive documents would have little effect, but at a similar richness, review effort could increase by an order of magnitude.

On the other hand, aiR prompt engineering and S-CAL training are sunk costs, independent of collection size or the number of responsive documents. We have no idea of the cost to run aiR over 1.4M documents; S-CAL can be trained and run on commodity hardware in 8 seconds elapsed time and 32 seconds CPU time, once the documents have been vectorized, which takes less than one hour.

Keeling et al. characterize 7.3% responsive as “low richness (7%, well below the 20–25% design target)” and suggest that with higher richness the methods tested would have achieved superior precision. While we find no guidance or empirical result supporting these assertions, we observe that, if they were true, the converse would hold: The methods tested would achieve an inferior result on the full corpus because of its lower richness. We note that the richness of the TREC 2009 Legal track topics, which were used to establish the effectiveness and efficiency of TAR [7], ranged in richness from 0.3% to 1.5% (median 0.6%) and achieved precision ranging from 67.3% (with 86.5% recall) to 88.4% (with 69.2% recall), in a collection of 569,034 emails.

For this particular task, the Redgrave Case Study shows aiR to be comparable to, but not superior to RAL, whether the tool or the L1 reviewer makes the final determination of responsiveness. When the L1 reviewer makes the final determination, CAL achieves the same recall as RAL, with less review effort. When the tool makes the final determination, S-CAL, trained on 2,000 L1 reviewer responsiveness determinations, is as effective as aiR, offers a calibrated adjustment of the recall-precision trade-off, and runs quickly and cheaply on commodity hardware. aiR offers only two viable recall-precision choices, offers no calibrated means to choose, and runs slowly on specialized hardware.

## 8 Acknowledgement

We thank Redgrave LLP for making the case-study data available for independent research and for their assistance in understanding the data and study methodology.

## Bibliography

- [1] Industry Documents Library - Opioids Industry - More Resources - About OIDA. <https://www.industrydocuments.ucsf.edu/opioids/more-resources/about-oida/>
- [2] Cormack, G.V., Grossman, M.R.: Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. pp. 153–162. SIGIR '14, Association for Computing Machinery, New York, NY, USA (Jul 2014). <https://doi.org/10.1145/2600428.2609601>
- [3] Cormack, G.V., Grossman, M.R.: Scalability of Continuous Active Learning for Reliable High-Recall Text Classification. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management. pp. 1039–1048. CIKM '16, Association for Computing Machinery, New York, NY, USA (Oct 2016). <https://doi.org/10.1145/2983323.2983776>
- [4] Cormack, G.V., Grossman, M.R.: Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 5–14. SIGIR '17, Association for Computing Machinery, New York, NY, USA (Aug 2017). <https://doi.org/10.1145/3077136.3080812>
- [5] Cormack, G.V., Grossman, M.R., Harbison, A., O'Halloran, T., McManus, B.: Unbiased Validation of Technology-Assisted Review for eDiscovery. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2677–2681. SIGIR '24, Association for Computing Machinery, New York, NY, USA (Jul 2024). <https://doi.org/10.1145/3626772.3657903>
- [6] Ghelani, N., Cormack, G.V., Smucker, M.D.: Refresh strategies in continuous active learning. In: Proceedings of the First International Workshop on Professional Search (ProfS 2018). CEUR Workshop Proceedings, vol. 2127, pp. 18–23 (2018)
- [7] Grossman, M., Cormack, G.: Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law & Technology* **17**(3), 11 (Jan 2011)
- [8] Grossman, M.R., Cormack, G.V., Roegiest, A.: TREC 2016 total recall track overview. In: Proceedings of the Twenty-Fifth Text Retrieval Conference (TREC 2016). No. 500-321 in NIST Special Publication, National Institute of Standards and Technology (2016)
- [9] Keeling, R.D., Mangum, R., Nelson, F.E., Reiss, K.A.: Generative AI for complex document review: A comparative evaluation benchmarked against active learning and an independent expert reviewer. Working Paper 2026-01, Redgrave LLP (Jun 2026)

- [10] Roegiest, A., Cormack, G.V., Grossman, M.R., Clarke, C.L.A.: TREC 2015 total recall track overview. In: Proceedings of the Twenty-Fourth Text Retrieval Conference (TREC 2015). NIST Special Publication, National Institute of Standards and Technology (2015)
- [11] Roitblat, H.L., Kershaw, A., Oot, P.: Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology* **61**(1), 70–80 (2010). <https://doi.org/10.1002/asi.21233>
- [12] Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management* **36**(5), 697–716 (Sep 2000). [https://doi.org/10.1016/S0306-4573\(00\)00010-8](https://doi.org/10.1016/S0306-4573(00)00010-8)
- [13] Webber, W., Chandar, P., Carterette, B.: Alternative assessor disagreement and retrieval depth. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management. pp. 125–134. CIKM '12, Association for Computing Machinery, New York, NY, USA (Oct 2012). <https://doi.org/10.1145/2396761.2396781>