PROMPTING CHANGE WITH GENAI IN LARGE-SCALE DOCUMENT REVIEW: A REAL-WORLD STUDY

Robert Keeling,[*] *Ray Mangum*,[**] *Amy Hanke*[***] & *Alyssa Ogden*[****]

INTRODUCTION

The evolution of document review in civil litigation has reached another milestone—one that invites a new partnership between humans and technology that promises gains in accuracy, efficiency, and scalability. Generative artificial intelligence (genAI) is remaking business and professional life at extraordinary speed,[1] and, within the legal sphere, nowhere is its impact more immediate—or the potential changes more dramatic—than document review.

Yet significant questions remain. When does genAI meaningfully improve document review? How should it be integrated into established workflows? How should its outputs be validated? As with prior technologies that are now widely accepted in eDiscovery, genAI's broader adoption will depend on demonstrated reliability and value supported by empirical data.

---

* Partner, Redgrave LLP.
** Partner, Redgrave LLP.
*** Counsel, Redgrave LLP.
**** Counsel, Redgrave LLP.

1. *See, e.g.*, Charlie Campbell et al., *The Architects of AI Are TIME's 2025 Person of the Year*, TIME MAG. (Dec. 11, 2025), https://time.com/7339685/person-of-the-year-2025-ai-architects/ [https://perma.cc/GX7T-MCLX] (noting the rapid advancement of generative AI technology).

Document review in civil litigation has undergone significant transformations before.[2] In the 1990s and earlier, review was largely manual: teams of attorneys examined thousands (and sometimes millions) of paper documents in warehouses over months. The early 2000s ushered in "e-discovery," as digital document collections and keyword searching became central to review and production. In the 2010s, technology-assisted review (TAR) (a.k.a. "predictive coding") introduced more sophisticated tools based on machine-learning—along with much debate about acceptable use, validation, and defensibility. With each evolution came novel issues, skepticism, and—ultimately—broad adoption.[3]

Today, the claim is that genAI can perform large-scale document review more effectively, economically, and defensibly than earlier approaches. The profession needs more empirical data to test that claim—to evaluate whether genAI truly delivers on its potential across varying types of legal matters and document sets. Among the key questions:

- Does genAI accurately and consistently predict relevant vs. non-relevant documents better than earlier technologies?
- Does genAI outperform or at least match human reviewers?
- Is genAI a defensible review methodology under legal discovery standards?
- Are there use cases where genAI does not perform well?
- What are the actual time and cost savings?
- How can lawyers best integrate genAI into existing workflows?
- What risks, if any, arise from using genAI in document review?

In this Article, we seek to answer some of these questions with empirical data, providing insight into the use of genAI to assist with document review in large-scale legal and regulatory matters.

For this case study, we worked with a client company and an eDiscovery vendor to select a 1,600-document sample from the client's prior, real-life legal matter. We then compared the relevance and issue coding of human reviewers to the coding decisions made by a market-leading genAI tool, Relativity's aiR for Review. The results varied depending on document type, but overall, the relevance review produced a high recall rate of 83.9% and a precision rate of 84.7%. The issue-coding results were more mixed, with accuracy often depending on the

2. *See, e.g.*, Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 11 (2011); George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10 (2007).

3. *See generally* Charlie Hernandez, *Tech Shifts & the Law*, 48-AUG L.A. LAW 26 (July/Aug. 2025).

nuances of the specific issue. The testing also revealed certain limitations of the tool. Based on these results, this Article concludes with recommendations for potential genAI use cases within a typical document review workflow and discusses the potential cost savings, including the factors that may affect that calculus.

Part I of this Article describes the evolution of document review methodologies from manual review to TAR and concludes with the rise of genAI and its early impact on the legal profession. In Part II, we discuss the need for empirical data to evaluate the utility and defensibility of genAI tools in document review, and we describe our proof-of-concept study. In Part III, we offer analysis of the study's results and provide our observations, learnings, and recommendations on how lawyers and their clients can best leverage genAI to enhance efficiency, improve accuracy, and make informed decisions about integrating those technologies into their existing document review workflows.

I. THE EVOLUTION OF TECHNOLOGY IN LAW MARCHES ON WITH GENAI

The evolution of technology in law provides a useful starting point for understanding both the hesitation and the promise surrounding state-of-the-art genAI tools in legal document review. Each technological advancement in the progression—from digital keyword searching to TAR—has sparked a mix of optimism and skepticism within the legal community. Technology adoption in law has historically been slow and steady, and the past serves as a helpful guide to what the inevitable adoption and integration of genAI in eDiscovery may look like.[4]

The origins of discovery in civil litigation began with manual, paper-based document review—an arduous and labor-intensive endeavor that was often inefficient, costly, and highly prone to human error.[5] This approach, now largely obsolete, is chiefly remembered by attorneys who practiced before the digital transformation of the twenty-first century.[6]

---

4. *See* Hernandez, *supra* note 3, at 28 (tracing the historical patterns of technology adoption in law and concluding that it demonstrates a predictable cycle of "initial fear and resistance, followed by increasing client pressure, eventual regulatory guidance, and ultimately, widespread acceptance").

5. *See* Dana A. Remus, *The Uncertain Promise of Predictive Coding*, 99 IOWA L. REV. 1691, 1702–03 (2014) (citing Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 1, 3 (2011)); Maura R. Grossman & Gordon V. Cormack, *Inconsistent Responsiveness Determination in Document Review: Difference of Opinion or Human Error?*, 32 PACE L. REV. 267 (2012)); The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 199 (2007).

6. *See* Bennett B. Borden & Jason R. Baron, *Finding the Signal in the Noise: Information Governance, Analytics, and the Future of Legal Practice*, 20 RICH. J.L. & TECH. 7, 3 (2014) (describing "the beginning" as manual review, a process of "legions of lawyers with hundreds if

By the early 2000s, the digital transformation was well underway—organizations across industries had adopted electronic solutions for record-keeping and data storage (e.g., cloud computing), communication (e.g., email), and transacting business (e.g., electronic contracts and e-signatures).[7] The result was an explosion of electronically stored information (ESI) that made continued reliance on traditional manual document review impractical. Despite uncertainty and concerns about the "black box" of technology, this unprecedented growth in ESI compelled the legal community to evolve its thinking and practices.[8]

Regulatory and judicial guidance lagged behind the rapid shift to digital business processes. Between 2003–2004, the landmark decisions in *Zubulake v. UBS Warburg LLC* established the foundation of modern-day ESI preservation obligations by recognizing that parties have certain duties to preserve and produce relevant ESI.[9] The court also addressed concerns that the overwhelming volume of ESI could render litigation cost-prohibitive by offering guidance on proportionality and cost-sharing principles for handling large-scale electronic data production.[10] Shortly thereafter, the 2006 amendments to the *Federal Rules of Civil Procedure* formally recognized ESI as a discoverable category of evidence and established rules specifically addressing electronic discovery and some of the challenges it raised.[11] Moreover, in 2012, the American Bar Association revised Model Rule 1.1 related to the duty of competence to expressly recognize the role of eDiscovery, requiring that lawyers stay

---

not thousands of boxes in warehouses, reviewing folders and pages one-by-one in an effort to find the relevant needles in the haystack").

7. Hernandez, *supra* note 3, at 28–31.

8. Hernandez, *supra* note 3, at 31 (noting that "law firms were forced to play ball" following the "rapid digitization" of the early 2000s); Kate Bauer, *Technology-Assisted Review: Overcoming the Judicial Double-Standard*, RICH. J.L. & TECH. BLOG (Jan. 24, 2018), https://jolt.richmond.edu/2018/01/24/technology-assisted-review-overcoming-the-judicial-double-standard/ [https://perma.cc/T3AB-8RGY] (calling for greater acceptance of TAR over manual review in light of "increasing document volumes and research on the shortcomings of traditional review methods"); Paul E. Burns & Mindy M. Morton, *Technology-Assisted Review: The Judicial Pioneers*, THE SEDONA CONF. INST. (Mar. 2014), https://www.americanbar.org/content/dam/aba/publications/litigation_committees/commercial/materials/technology-assisted-review-the-judicial-pioneers.pdf [https://perma.cc/3KWN-ZMHL] (noting plaintiffs' objections in the *Da Silva Moore* case that they do not understand the "black box" of predictive coding and "there is no way to be certain if MSL's method is reliable").

9. Zubulake v. UBS Warburg LLC, 229 F.R.D. 422, 431 (S.D.N.Y. 2004); *but see* Robert Keeling, *Sometimes, Old Rules Know Best: Returning to Common Law Conceptions of the Duty to Preserve in the Digital Information Age*, 67 CATH. U. L. REV. 67, 102 (2018) (indicating that there is no duty under traditional common law to preserve until a lawsuit is filed or imminent).

10. *See* Zubulake v. UBS Warburg LLC, 217 F.R.D. 309, 317–18 (S.D.N.Y. 2003).

11. *See* Burke T. Ward et al., *Electronic Discovery: Rules for a Digital Age*, 18 B.U. J. SCI. & TECH. L. 150, 179 (2012).

apprised of "the benefits and risks associated with relevant technology."[12] At this point, using electronic solutions such as keyword searches and applying metadata filters to a document set had become the norm.[13]

But as lawyers and their teams of document reviewers struggled to keep up with the ever-growing volumes of ESI, new technology solutions using predictive coding emerged.[14] Computer-assisted review based on supervised machine learning—at the time often called "predictive coding" and today more commonly referred to as technology-assisted review (TAR)—is a process whereby machine learning algorithms are trained by human reviewers to predict the relevance of documents within large data sets and then applied to entire document populations to classify documents, improving both speed and accuracy over manual review and keyword searching.[15]

As before, the practical application of TAR outpaced the development of judicial and regulatory guidance. In 2012, the landmark case of *Da Silva Moore v. Publicis Groupe* became the first federal court decision to approve the use of TAR in document review.[16] In support of his holding that a producing party could use TAR in appropriate cases, Magistrate Judge Peck explained the challenges of manual review and keyword searching, examined the empirical data in support of integrating TAR, and concluded that TAR was better than the alternatives in that case.[17] The defensible use of TAR in document review was fortified in 2015 by *Rio Tinto PLC v. Vale S.A.*, which confirmed as black letter law that TAR is an acceptable discovery tool and that a producing party does not need prior approval from the opposing party or the court to implement it.[18]

---

12. MODEL RULES OF PRO. CONDUCT r. 1.1 cmt. 8 (A.B.A. 2020); A.B.A. Comm'n on Ethics 20/20, *Resolution and Report on Technology and Confidentiality 105A* 3 (Aug. 2012); *see also* Lori D. Johnson, *Navigating Technology Competence in Transactional Practice*, 65 VILL. L. REV. 159, 168 (2020) (explaining the significance of the ABA's adoption of amended Model Rule 1.1 regarding the duty of competence to include technological competence).

13. *See* The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 200 (2007) ("By far the most commonly used search methodology today is the use of 'keyword searches' of full text and metadata as a means of filtering data for producing responsive documents in civil discovery.").

14. *See* Robert Keeling et al., *Using Machine Learning on Legal Matters: Paying Attention to the Data Behind the Curtain*, 11 HASTINGS SCI. & TECH. L.J. 9 (2020); *see also* Grossman & Cormack, *supra* note 2, at ¶ 28 (offering evidence that technology-assisted review yields superior results to manual review); *see also* Robert Keeling et al., *Separating the Privileged Wheat from the Chaff–Using Text Analytics and Machine Learning to Protect Attorney-Client Privilege*, 25 RICH. J.L. & TECH. 2, ¶¶ 33–34 (2019).

15. *See* Charles Yablon & Nick Landsman-Roos, *Predictive Coding: Emerging Questions and Concerns*, 64 S.C. L. REV. 633, 634 (2013).

16. 287 F.R.D. 182, 193 (S.D.N.Y. 2012).

17. *Id.* at 189–91. *See also* Bauer, *supra* note 8.

18. Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 126–27 (S.D.N.Y. 2015).

Critically, *Rio Tinto* further confirmed that TAR should not be held to a higher standard than other document review methodologies, such as manual review or keyword searching.[19]

Over the next decade TAR became a routine part of large-scale document review workflows, with widespread adoption by law firms and corporate legal departments as both courts and clients recognized its increased efficiency, accuracy, and cost-effectiveness in comparison to prior methods.[20] Courts continued to provide guidance as new questions arose in the context of applying TAR, such as the responding party's discretion to use TAR, the effect of ESI protocols on a party's decision to use TAR, transparency and disclosure requirements, TAR methodologies and workflows, proportionality, and cost-shifting.[21] For example, following the acceptance of TAR, a line of cases held that the producing party typically has discretion to decide whether to use TAR and which TAR methodology to employ, as long as the process is reasonable.[22] Subsequent caselaw grappled with technical aspects of using TAR, as well as the use of TAR in conjunction with other review methodologies.[23]

Each stage of technological evolution in discovery has been marked by an increase in the volume and complexity of data, the straining of existing review methodologies, skepticism about the emerging technology, validation of that technology through experience and studies, acceptance within the judiciary and legal community, and, finally, normalization of the technology into standard workflows.[24] GenAI presents the next inflection point. The unrelenting growth of ESI and the advent of new data sources are fueling this continued evolution. With genAI, companies and their counsel are seeking more sophisticated eDiscovery tools to manage overwhelming volumes of diverse and

---

19. *Id.* at 129.

20. *See generally* The Sedona Conference, *The Sedona Conference TAR Case Law Primer, Second Edition*, 24 SEDONA CONF. J. 1 (2023) [hereinafter *TAR Case Law Primer*].

21. *Id.* (citing cases).

22. *See id.* at 27–29 (citing Livingston v. City of Chicago, No. 16 CV 10156, 2020 WL 5253848 at *3 (N.D. Ill. Sep. 3, 2020); Coventry Cap. US LLC v. EEA Life Settlements Inc., No. 17-Civ. 7417 (VM) (SLC), 2020 WL 7383940, at *4 (S.D.N.Y. Dec. 16, 2020), *objections overruled*, 2021 WL 961750 (S.D.N.Y. Mar. 15, 2021); Lawson v. Spirit AeroSystems, Inc., No. 18-1100-EFM-ADM, 2020 WL 1813395, at *8–9 (D. Kan. Apr. 9, 2020); Kaye v. N.Y.C. Health and Hospitals Corp., No. 18-CV-12137 (JPO) (JLC), 2020 WL 283702 (S.D.N.Y. Jan. 21, 2020)).

23. *TAR Case Law Primer*, *supra* note 20, at 48–62.

24. *See, e.g.*, Hernandez, *supra* note 3, at 31–32 (analyzing the historical trajectory of technology adoption in law: "necessity ultimately drives acceptance, with judicial clarification and regulatory adaptation cementing these innovations into standard legal practice"); Matthew G. Kenney, *The Past, Present and Future of Predictive Coding*, 12 FLA. A&M U.L. REV. 165, 176–78 (2016) (analyzing the "slow adoption rate" of predictive coding in document review by lawyers even after judicial acceptance in *Da Silva Moore*).

complex data under tight deadlines in large-scale litigation and regulatory investigations.

The results of a recent study confirm that in-house legal teams see the potential benefits of genAI and are driving adoption at a rapid pace.[25] According to a 2025 survey by the Association of Corporate Counsel of 657 in-house legal professionals from 30 countries, the shift from "passive planning to active implementation" of genAI is underway and moving swiftly to embrace the integration of genAI into legal matters.[26] The 2025 survey shows that genAI adoption by U.S. legal departments has more than doubled since 2024, with 52% of respondents reporting that they are already using genAI in their legal practice.[27] Only 2% of respondents were neither using nor planning to use genAI in 2025.[28] In terms of its perceived benefits, 91% of respondents stated that efficiency was the primary benefit of incorporating genAI into their legal matters.[29] Notably, however, the study found that in-house legal departments are not yet seeing the cost-savings benefit of genAI from outside counsel.[30]

While genAI promises even greater efficiency and cost savings than the technology solutions that preceded it, there is work to be done within the legal community to ensure its acceptance and successful integration in eDiscovery. Empirical research will be useful in further demonstrating its effectiveness.[31] Legal uncertainty adds another layer of complexity: courts and regulatory bodies are only beginning to address AI-generated outputs. The Sedona Conference and other legal organizations have begun issuing guidance on responsible AI use in eDiscovery, highlighting best practices.[32]

As the legal community evaluates genAI, its acceptance will likely depend on the availability of empirical research, judicial guidance, and standards to address new risks. Unlike prior tools that classified

---

25. ASS'N OF CORP. COUNS., *Generative AI's Growing Strategic Value for Corporate Law Departments – Survey Results* (Oct. 14, 2025), https://www.acc.com/resource-library/generative-ais-growing-strategic-value-corporate-law-departments-survey-results [https://perma.cc/2NNF-KTGZ].

26. *Id.*

27. *Id.*

28. *Id.*

29. *Id.*

30. *Id.*

31. *See* Maura R. Grossman et al., *Does the LLMperor Have New Clothes? Some Thoughts on the Use of LLMs in eDiscovery*, NAT'L L. REV. (Nov. 4, 2024), https://natlawreview.com/article/does-llmperor-have-new-clothes-some-thoughts-use-llms-ediscovery [https://perma.cc/V2QY-EXHY] (concluding that empirical studies are necessary to demonstrate the effectiveness of LLM tools in eDiscovery).

32. *See* The Sedona Conference, *Primer on Generative AI in Discovery, Draft* (2025); The Sedona Conference, *The Sedona Canada Primer on Artificial Intelligence and the Practice of Law*, 26 SEDONA CONF. J. 99 (2025); Judge Xavier Rodriguez, *Artificial Intelligence (AI) and the Practice of Law*, 24 SEDONA CONF. J. 783 (2023).

documents based on large amounts of training data, genAI tools allow lawyers to describe what they are looking for in natural language. Moreover, many genAI tools provide narrative explanations and other outputs to support coding decisions, which allows reviewers to understand and evaluate the basis for the decision. The distinctions between these tools raise new questions about reliability and efficiency that cannot be answered simply by analogy to TAR and the body of evidence supporting its use. To that end, this Article presents a study that provides empirical data to help evaluate the use of genAI tools in large-scale document review, including an evaluation of their accuracy, efficiency, and limitations.

II. DATA-DRIVEN GENAI ADOPTION: INSIGHTS FROM AN INNOVATIVE LEGAL STUDY

A key aspect of the adoption of prior technologies in eDiscovery has been the existence of empirical studies to support the effective and defensible use of the emerging technologies. The primary example is Magistrate Judge Peck's seminal 2012 *Da Silva Moore* decision, which expressly approved of using computer-assisted review methodologies in discovery in appropriate cases and made such acceptable use black letter law.[33] This judicial endorsement was a big step in establishing predictability for litigants and lawyers that the use of TAR and similar technologies would be legally defensible, which helped to solidify integration of the technology as a routine part of eDiscovery plans for large-scale document review.[34]

A significant factor in Judge Peck's endorsement of TAR was the existence of empirical data showing an increase in accuracy and the cost-saving benefits of using TAR over manual human review in voluminous cases.[35] Based on his review of several data-driven studies, Judge Peck concluded: "Computer-assisted review appears to be better than the available alternatives, and thus should be used in appropriate cases."[36]

Now, as market demands and ever-growing data volumes fuel the push for rapid adoption of genAI technology in the law, there is a need for more data-driven research to support the defensibility and judicial acceptance of using genAI tools, as well as to guide the responsible integration of genAI-based technologies into document review

33. Da Silva Moore v. Publicis Groupe, 287 F.R.D. 182, 183 (S.D.N.Y. 2012)*; see also* Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 127 (S.D.N.Y. 2015) (citing *Moore* and subsequent caselaw for the proposition that the acceptance of TAR is black letter law).

34. *Moore*, 287 F.R.D. at 193 ("Counsel no longer have to worry about being the 'first' or 'guinea pig' for judicial acceptance of computer-assisted review.").

35. *Id.* at 190–91 (citing studies).

36. *Id.* at 191.

workflows.[37] This type of empirical support will help build the foundation necessary to demonstrate the effectiveness, accuracy, and cost-efficiency of genAI, fostering greater trust among clients, lawyers, and the judiciary, and encouraging broader acceptance in the legal community.

Given the relative infancy of genAI legal tools, there is currently little empirical data evaluating genAI in the law.[38] To help fill this void, the authors[39] conducted a study that incorporated actual client data from a prior legal matter to evaluate the performance of a genAI tool to conduct first-level document review for responsiveness and issue coding in a large-scale document review. The study results highlight the capabilities of genAI, as well as identify some challenges and limitations. The study offers valuable guidance for the legal community by assessing the effectiveness of genAI tools in document review, outlining best practices for their integration into current workflows, and providing insights that inform future use cases and workflow strategies.

A. *The Study Protocol*

In our practice, the authors have leveraged cutting-edge genAI tools to assist with document review in several active client matters. The genAI review tools are powered by large language models and simulate human document review. Use cases are proliferating, but common workflows include relevance review, issue review, privilege review, and key document identification. For each analysis type, the user enters prompts analogous to the review protocol in a traditional review workflow. The tool then analyzes the extracted text of the documents according to the prompt and provides predictions. Many genAI tools also provide the rationale for its score prediction and citations within the document that support the prediction. The results can be used in a variety of ways, ranging from replacing first-level review, identifying documents to prioritize for human review, or conducting quality control of human reviewers.

---

37. *See* Grossman et al., *supra* note 31 (concluding that empirical studies are necessary to demonstrate the effectiveness of LLM tools in eDiscovery).

38. *Id.* ("As far as we are aware, the impact of this phenomenon on eDiscovery search has neither been researched nor reported. . . . No study has yet shown either approach to be superior to state-of-the-art TAR methods."). *See also* Eugene Yang et al., *Beyond the Bar: Generative AI as a Transformative Component in Legal Document Review*, IEEE INT'L CONF. ON BIG DATA (Dec. 2024) (presenting empirical data from a legal matter evaluating an LLM-based document review system, reporting 96% recall and 60% precision without matter-specific tuning).

39. Redgrave LLP is a law firm specializing in information law, including eDiscovery, information governance, AI governance, data privacy, and cybersecurity. *Our Practice*, REDGRAVE LLP, https://www.redgravellp.com/our-practice [https://perma.cc/KVB7-65UZ] (last visited Mar. 1, 2026).

Working with an eDiscovery vendor, the authors initiated a study to assess the real-world performance of genAI tools. The study protocol was designed to evaluate genAI's ability to code documents for relevance and issue coding in an apples-to-apples comparison against human reviewers, considering both accuracy of coding decisions and efficiency from a cost and time perspective.

To facilitate this study, the authors worked with a corporate client to identify a prior closed matter in which the company was responding to a subpoena from a regulatory agency. During the actual review, attorney reviewers coded the document set for responsiveness and applied issue coding that correlated with the specific requests outlined in the subpoena. The document population consisted of custodial data (primarily emails and attachments), as well as documents collected from certain workpaper documents (including Word documents and Excel spreadsheets). Because of the nature of the subject matter of the workpaper documents, cover pages were created during collection of those documents for the purpose of providing context to the attorney reviewers in the underlying legal matter. The workpaper documents and associated cover pages were structured as a family during the human review.

The first step in the study protocol called for segregation from the review population of any documents that were deemed ineligible by the genAI tool. Like search terms and TAR, the genAI tool here works on the extracted text of a document. Thus, documents that do not contain extracted text (e.g., images or container files) and documents where the extracted text is too small or too large are ineligible to be processed through the tool. The study protocol excluded any documents without extracted text or with an extracted text size of under .05 KB or over 150 KB, resulting in approximately 6% of the documents being excluded as ineligible. Within the document population, the extracted text restrictions had little impact on emails and common attachment files, but they did eliminate several other files, including workpapers.

Once ineligible files were removed, the study protocol called for identification of statistical samples of documents previously coded by attorney reviewers, including a sample of 1,600 custodial documents, stratified by custodian and document type. The sample set represented all the possible issue codes, as well as a mix of responsive and nonresponsive coding. A similar, separate sample of 200 previously coded documents was identified for the population of workpapers.

Next, Redgrave attorneys drafted prompts to guide the responsiveness review. These initial prompts were based on the same information provided to attorney reviewers in the original document review protocol

in the underlying matter, with only minor revisions.[40] The study team then worked to iteratively refine the prompt by evaluating its performance on small samples of fifty documents, making revisions based on the results of each round.[41]

B. *GenAI Categorization*

The genAI tool used in this study (Relativity's aiR for Review) reports both a prediction and a numerical score for each analyzed document. The prediction is the relevance, key-document, or issue label aiR recommends; the score reflects how strongly relevant the document is, or how well it matches an issue. More specifically, the genAI tool scores documents from 0 to 4 according to their predicted level of relevance.[42] A score of 4 ("Highly Relevant") indicates direct, strong evidence that the document relates to the case or issue. A score of 3 ("Relevant") indicates the document is predicted relevant with citations to supporting text. A score of 2 ("Borderline Relevant") indicates that aiR found some content that might relate to the case or issue. A score of 1 ("Not Relevant") indicates aiR did not find evidence of relevance. A score of 0 ("Junk" or "Highly Not Relevant") indicates the document contains no useful information, such as system files or empty documents. A score of -1 indicates an error—the document either encountered an error or could not be analyzed. A table from Relativity's website summarizes its methodology:

Score	**Description**
4	**Highly Relevant**: The document is predicted very relevant to the issue. aiR found direct, strong evidence that the content relates to the case or issue. Citations show the relevant text.
3	**Relevant:** The document is predicted relevant to the issue. Citations show the relevant text.
2	**Borderline Relevant:** The document is predicted borderline relevant. aiR found some content that might relate to the case or issue. It usually has citations.

40. The genAI tool's (i.e., aiR for Review) prompt template for Responsiveness Analysis includes text boxes related to the following: matter overview, relevant people and aliases, noteworthy organizations, noteworthy terms, additional context, and responsiveness criteria.

41. During the time of the study, the authors understood that Relativity aiR products were using GPT-4o May 2024. Run time for sets of 50 documents was approximately 3 minutes, for both Relevance and Issue Analysis. Run time for the set of 1,600 was approximately 35 minutes.

42. *See Understanding Document Scores*, RELATIVITY https://help.relativity.com/RelativityOne/Content/Relativity/aiR_for_Review/aiR_for_Review_results.htm#Understanding_document_scores [https://perma.cc/EFW3-9B9Z] (last visited June 13, 2026).

1	**Not Relevant:** The document is predicted not relevant. aiR did not find any evidence that it relates to the case or issue.
0	**Junk**: The document contains no useful information or is considered "junk" data, such as system files, an empty document, or sets of random characters.
-1	**Error**: The document either encountered an error or could not be analyzed. For more information, see How document errors are handled.

At this point, two review project managers (the "conflict reviewers") from the underlying matter reviewed conflicts between the attorney coding and the genAI predictions, providing a final call on responsiveness and an explanation for the reasoning. Where the conflict review resulted in overturning the genAI tool's prediction, the study team discussed both the rationale provided by the genAI tool and the explanation provided by the conflict reviewers to determine whether the prompts should be revised.

In the first set of fifty documents, there were six conflicts, five of which were resolved in favor of the genAI tool. There were also eight documents predicted as borderline responsive, all of which were determined to be nonresponsive. The study team determined that iteration was required to address over-reliance on certain industry-based terms as indicia of responsiveness to a specific request in the subpoena. This was accomplished by revising the syntax of the relevant sentence in the prompts and adding to the relevance criteria that certain industry-specific concepts were not independently relevant.

The revised prompt was then run on a second set of fifty documents. On the second set, there were ten conflicts, four of which were resolved in favor of the genAI tool, and eight borderline documents. Following discussion of the conflicts, the study team determined that the prompt should be revised once more to address over-reliance on references to a single industry-specific keyword as indicia of responsiveness to the subpoena requests.

Once the study team determined that the prompt was sufficient, the team ran the genAI tool on the control set of 1,600 custodial documents. To ensure the reliability of the gold standard determinations, the study team implemented a rigorous validation process. Two attorneys who had led the original document review in the underlying legal matter served as conflict reviewers and reviewed the full 1,600-document sample.[43] These reviewers were given the review protocol and Q&A log from the original matter and were instructed to code each document for relevance without

43. The two conflict reviewers split the 1,600-document set between them without duplication in the review.

deference to either the aiR predictions or the prior attorney coding. Each reviewer provided written notes explaining the basis for coding decisions on each document. The resulting validation effort covered all 1,600 documents in the sample and provided the final human coding used as the "gold standard" for the responsiveness metrics reported below.

In addition to identifying responsive documents, genAI can also be used for issue analysis, categorizing documents based on a set of defined legal issues to allow for more in-depth review and case assessment. To test the efficacy of this use case, the refined prompt was adapted for use in the issue tagging analysis and run on the 1,600 custodial document sample, as well as a sample of fifty of the workpapers.[44] The conflict reviewers then reviewed conflicts from the custodial document sample, but as discussed below, following initial review of the results, the study team decided that a full conflict review of the workpapers for issue analysis was not warranted.

III. STUDY RESULTS SUPPORT INCORPORATION OF GENAI IN RESPONSIVENESS REVIEW

GenAI's application in legal document review is still in its early stages, and important questions remain about its efficacy and optimal use cases. This study provides meaningful data on several of those questions, though additional research will be needed as the technology and its applications continue to evolve. Our findings offer a foundation for informed conversation and a starting point for the continued empirical work that will shape genAI's role in eDiscovery.

A. *Responsiveness Analysis for Custodial Documents*

The results of our study demonstrate that genAI is an effective and reliable tool for use in responsiveness review workflows for electronic documents typically found in custodial collections (e.g., email and common attachments like Microsoft Word documents). With respect to responsiveness review, genAI returned more-than-adequate recall and very good precision, identifying actually responsive documents on par with attorney reviewers while generating fewer false positives than attorney reviewers. It also exceeded the widely accepted standards of 70-75% recall for TAR models.[45]

---

44. The Issue Analysis prompt template is the same as the responsiveness review except that the general "Responsiveness Criteria" text box is replaced with up to ten responsiveness criteria for particular issue tags. Here, we used the ten most common issue tags selected by attorney reviewers in the underlying legal matter.

45. Recall is the percentage of all relevant documents identified by the model, whereas precision is the percentage of documents the model indicated were relevant that actually were relevant. Recall and precision are usually inversely proportionate measures, meaning that as recall increases, precision usually decreases. Keeling, *supra* note 14, at 16–17. A commonly accepted

In the control set of 1,600 custodial documents, genAI predicted that 554 were responsive.[46] The gold standard review, conducted by two attorneys who led the original review and who reviewed the full 1,600-document sample, determined that 559 documents were in fact responsive.

Comparing genAI's predictions to the "gold standard" determinations, genAI achieved strong performance on definitive predictions. Of the 554 documents predicted responsive, 469 (84.7%) were confirmed responsive (true positives), while 85 (15.3%) were not responsive (false positives). Of the 600 documents predicted not responsive, 590 (98.3%) were confirmed not responsive (true negatives), while 10 (1.7%) were responsive (false negatives).

The gold standard reviewers also evaluated the 368 documents predicted as borderline responsive by genAI. Of those, 69 were determined responsive by the gold standard review, while the remaining 299 were determined not responsive. Additionally, genAI classified 62 documents as junk, of which 2 were determined responsive. Of the 16 documents for which genAI returned an error, 9 were determined responsive by the gold standard review.

As a result, excluding borderline documents from the analysis, the calculated recall rate was 83.9% (469/559) and the precision rate was 84.7% (469/554). In other words, genAI correctly identified approximately 84 out of every 100 responsive documents and, of the documents predicted by the genAI tool as responsive, approximately 85 out of 100 were in fact responsive. When borderline documents were included as responsive predictions, recall increased to 96.2% ((469+69)/559) but precision dropped to 58.4% ((469+69)/(554+368)). A table summarizing the results is below:

---

recall rate is 70–75%. Given the tradeoff between precision and recall, an acceptable precision rate typically depends on the needs of the case. *See* Neel Guha et al., *Vulnerabilities in Discovery Tech*, 35 HARV. J.L. & TECH. 581, 599–600 (2022) (citing Maura R. Grossman & Gordon V. Cormack, *Vetting and Validation of AI-Enabled Tools for Electronic Discovery*, *in* LITIGATING ARTIFICIAL INTELLIGENCE 407, 409 (Jill Presser, et al. Beatson & Gerald Chan eds., 2021) (discussing requirements on evaluation protocols)); *see also* Order Regarding Search Methodology for Electronically Stored Information at *6, *In re* Broiler Chicken Antitrust Litig., No. 16-cv-08637, 2018 WL 1146371 (N.D. Ill. Jan. 3, 2018).

46. All such documents had an "aiR Score" of 3. This score represents the genAI tool's confidence level on a scale of 0–4. Thus, all were predicted Relevant but none were predicted Highly Relevant.

Without Borderline		
Recall	83.9%	469/559 responsive documents identified
Precision	84.7%	469/554 predicted responsive documents actually responsive
With Borderline Documents Treated as Responsive		
Recall	96.2%	(469+69)/559 responsive documents identified
Precision	58.4%	(469+69)/(554+368) predicted responsive documents actually responsive

The baseline recall of 83.9% compares favorably to state-of-the-art performance metrics for traditional TAR models where a 70–75% recall rate is widely considered acceptable.[47] Precision of 84.7% for a TAR model would be considered highly effective at 70–75% recall and is especially impressive at 83.9% recall.

These results indicate that genAI can be a highly effective tool for identifying responsive documents within certain limitations of the tool. In addition to the text size limitation, which impacted approximately 6% of the document population, there are also limitations to the type of information genAI can consider. For example, most genAI tools (including the one used in this study) currently do not consider family relationships, the relationship between hyperlinked documents, or document metadata. Like TAR, the tool only reviews extracted text and therefore may not consider context in charts or graphs and may not consider text in images or handwriting if not accurately processed as text. In addition, the results show how including the so-called "borderline" documents can have a significant impact on efficacy and efficiency. Including the borderline documents can increase the number of responsive documents identified, while potentially materially increasing the number of false positives and thus increasing the cost of the review. A full summary of the data can be found below:

47. *See supra* note 45.

Table 1. Results of Custodial Documents Responsiveness Review Analysis

Custodial Documents Responsiveness Review	#	%
Sample population	**1600**	
Definitive predictions	1216	76.0%
Borderline predictions	368	23.0%
Errors[48]	16	1.0%
Definitive Population		
Disagreements between aiR and gold standard	95	7.8%
False positives (predicted R, gold standard NR)	85	15.3% of predicted R
False negatives (predicted NR, gold standard R)	10	1.7% of predicted NR
True positives (predicted R, gold standard R)	469	84.7% of predicted R
True negatives (predicted NR, gold standard NR)	590	98.3% of predicted NR
Junk documents (predicted junk, gold standard R)	2/62	3.2%
Borderline Population		
Determined responsive by gold standard	69/368	18.8%
Determined not responsive by gold standard	299/368	81.3%
Error Population		
Coded responsive by attorney reviewers	9/16	56.3%

48. On occasion, the genAI tool failed to return a result even when the document met the eligibility criteria. During execution, 26 documents errored out on the first run. Error details for 8 documents noted the "document text is too long" although none were over the suggested extracted text size. Details for one document listed "ungrounded citations detected in completion," which indicates that the genAI tool generated a citation that cannot be found in the document's text—suggesting a potential AI hallucination. Finally, 17 documents had unknown errors. Following the analysis, it was determined that running genAI a second time on the errored set eliminated 10 of the unknown errors. A third attempt did not eliminate any other errors.

Determined responsive by gold standard	9/16	56.3%
Determined not responsive by gold standard	7/16	43.8%
Overall Metrics		
Total true positives	469	
Predicted responsive and coded responsive	469	
Identified in conflict review	N/A	
Total false positives	85	
Total false negatives[49]	90	
Recall[50]	**469/559**	**83.9%**
Precision	**469/554**	**84.7%**

B. *Issue Analysis for Custodial Documents*

The results of genAI's issue analysis in the study were mixed and influenced by various factors, including document type, metadata fields, and the substantive issue being evaluated. For issue analysis, genAI was run on the same sample of 1,600 documents for the ten issue tags most often selected in the underlying review. Reviewers applied one or more of the ten issue tags to 468 of the 1,600 documents. As shown in Table 2 below, the rates at which genAI predicted that a document was relevant to an issue, as compared to human reviewers, varied across the issues.

The conflict reviewers reviewed the 380 documents for which genAI's prediction conflicted with attorney coding for one or more issue tags (23.8% of the sample). At a document level, conflicts for 167 documents (43.9%) were resolved fully in favor of genAI. Conflicts were resolved fully in favor of the reviewer for 126 documents (33.2%). For 87 documents (22.9%), there were multiple conflicts regarding multiple issues and the conflict reviewer agreed with some but not all of genAI's predictions.

49. Includes responsive documents predicted borderline, junk, or error.

50. Recall and precision are calculated assuming that borderline documents are excluded from production.

Table 2. Comparison of aiR Predictions to Reviewer Coding for Issue Tagging Analysis (Pre-Conflict Review)

	Issue # 1	**Issue # 2**	**Issue # 3**	**Issue # 4**	**Issue # 5**
aiR	93	247	100	122	23
Reviewer	27	363	90	220	57
Overlap	15	173	52	90	13

	Issue # 6	**Issue # 7**	**Issue # 8**	**Issue # 9**	**Issue # 10**
aiR	24	36	50	46	13
Reviewer	31	42	101	22	51
Overlap	19	23	47	7	9

At the issue level, post-conflict-review recall and precision estimates varied. Table 3 below contains estimates for three of the issues. Note that while recall and precision likely could have been improved for some of the issues through iteration, the results for other issues appear to reflect genAI's current limitations.

For example, Issue #1 called for information that was limited to a specific time period, namely the year 2023. The low precision and high recall rates for genAI's predictions for Issue #1 appear to be caused in large part by its inability to consider document metadata or other indicia of date. As a result, genAI predicted documents from other years were relevant, in addition to correctly identifying the relevant 2023 documents. In other words, it could not distinguish documents by year. By contrast, the documents responsive to Issue #5 (above) were largely spreadsheets and charts. GenAI's relatively low recall rate for Issue # 5 appears to be caused in part by a limited ability to identify the relevant information in this type of document format, for which relevance was more evident in native or image view than in extracted text.

At a high level, the issue analysis showed that where determinations were driven by the text of the document, genAI performed fairly well. Where the issue coding depended on metadata, structured content (e.g., spreadsheets or templates/forms), or documents without sufficient text (e.g., images), genAI's issue analysis performed much worse. It is possible that further iteration of the prompts for some issues could have improved the overall results. Another possible solution would be to apply metadata filters prior to running the documents through genAI where a metadata field is particularly relevant to the determination—such as where dates are critical to the determination.

Table 3. Recall and Precision Estimates for Example Issues

Issue Tag	Issue # 1	Issue # 3	Issue # 8
Predicted responsive by aiR	93	100	50
Tagged responsive by reviewer	27	90	101
True positives	17	64	50
aiR + reviewer agreement	15	52	47
Identified in conflict review	2	12	3
False positives	76	36	0
False negatives	0	8	11
Recall	**100%**	**64%**	**75%**
Precision	**18.2%**	**88%**	**100%**

C. *Analysis of Workpapers*

The workpapers collected in the underlying legal matter consisted largely of a specific type of document that is highly relevant in the client's industry. These documents included a number of characteristics that proved challenging when genAI was applied. After applying the issue-tagging prompts to a sample of the workpaper documents, the study team determined that genAI was not an effective tool for issue coding this particular type of industry-specific document at the time of the study. Accordingly, the analysis was not run on the full 200-document sample.

This assessment was helpful in highlighting several factors that may present limitations when using genAI tools to evaluate certain kinds of workpaper documents. First, because current genAI tools used for eDiscovery operate on the extracted text of a single document, genAI cannot consider family relationships or the context provided by family documents. In this study, this limitation was problematic because of the way data was stored and exported from the client's internal workpapers system. Specifically, during export, a cover page was created and assigned as the parent to the underlying document, which was assigned a child relationship. Because the cover page was processed as a separate document, the genAI tool did not have the benefit of context from the coversheets that were available to human reviewers when evaluating the document.

Similarly, most genAI tools currently cannot consider hyperlinked documents embedded within the underlying document. In this study, many of the workpapers contained hyperlinks to various support

documents, which provided additional context to the human reviewers that genAI could not consider at the time of the study.

Finally, several of the workpapers featured formatting issues that posed challenges for genAI. For example, many documents were templates—both completed and blank versions—which the tool processed as text files without recognizing the template formatting. As a result, genAI seemed to have difficulty distinguishing between actual content and template elements. Additionally, workpapers generally do not convert cleanly to extracted text, making it difficult for the tool to interpret their content accurately. Checklists were especially problematic, as checked boxes are not represented in the extracted text, further complicating genAI's ability to analyze the information accurately.

D. *Analysis of genAI's Cost Effectiveness*

Based on the study results showing the effectiveness of genAI in conducting responsiveness review, integrating genAI into a document review workflow in large-scale matters may represent significant cost savings in appropriate cases—particularly compared to manual review.

The cost savings would, of course, vary based on a number of document review-specific factors. For example, the percentage of the document population that is eligible for genAI will impact the calculus, with higher savings for document sets where the eligibility percentage is higher. Further, companies would see greater cost savings where there are fewer or no restrictions on the type of documents where genAI replaces first-level review—such as sensitive documents or C-suite custodians' documents. Cost savings would also be higher where second-level quality control by human reviewers is limited to a sample rather than the entire document set. Furthermore, the complexity of the document requests and the time spent on prompt iteration will also be a factor in the calculus on cost savings.

In addition, clients and their counsel should consider the potential for costly negotiations related to ESI protocols that call for the use of genAI, particularly in comparison to the well-established acceptance of TAR. Although integration of genAI may offer greater efficiency in conducting document reviews, its status as a novel technology not yet widely accepted by courts—and the potential for disputes with opposing counsel over matters such as transparency and validation—could lead to contentious and costly negotiations, potentially making it more expensive than traditional approaches at this point in time.

E. *Potential Use Cases and Future Workflow Considerations*

This study provides important empirical validation for the accuracy of genAI tools for inclusion in document review workflows. In particular, genAI seems well suited to conduct first-level responsiveness review of

text-based documents, subject to certain limitations. Although genAI tools may serve as a defensible and efficient component of the eDiscovery workflow in appropriate cases, human oversight and quality-control checks remain important to ensure accuracy and reliability. GenAI may also be effective as a means of prioritizing documents for attorney review or as a quality-control measure applied to certain document sets following attorney review. On the quality control front, the study results suggest that the tool may be particularly effective at identifying false positives—documents coded by attorneys as responsive that are not, in fact, responsive—thereby reducing the inadvertent production of nonresponsive documents. This study also provides empirical validation for the use of genAI tools for inclusion in document review workflows.

The study also offers valuable insight into the factors to consider when determining if genAI would be suitable for a particular legal matter and/or use case, as well as developing a proactive workflow that ensures the efficient use of the tool. A few practical considerations include:

- Extracted text: Consider the data set's characteristics to determine if genAI would add efficiency, keeping in mind that genAI tools work best on primarily text-based documents. Exclude from the genAI workstream any documents that do not meet the extracted text size specifications and other technical requirements.
- Family relationships: Factor in the inability of genAI tools to recognize family relationships and how that may impact its efficacy for evaluating certain documents. Consider whether workarounds may be feasible and economical. For example, to analyze the industry-specific workpaper documents in this study, the team could have created a version of extracted text that merged the cover page and document text for purposes of the genAI analysis, allowing genAI—like the human reviewers—to consider the context provided by the cover page when assessing responsiveness.
- Information beyond the four corners of a document: In some cases, information outside the four corners of the extracted text, such as metadata (e.g., date, document title, author, custodian, or file path) should be considered to make an accurate call for a category of documents. Consider excluding these documents from the genAI workstream or addressing them with a tailored workflow (e.g., modify the prompt prior to running that set of documents or run the documents through a metadata filter prior to analyzing with genAI).
- Accuracy of extracted text: The efficacy of genAI tools is dependent on the accuracy of the extracted text. In this study, for instance, documents containing checklists presented a particular challenge because the genAI tools used for this study cannot determine whether a box was checked. If the data set includes document categories that may have

incomplete or inaccurate extracted text, consider reviewing a sample of the extracted text to determine whether those documents should be excluded from the genAI workflow. Document categories susceptible to this limitation may include scanned documents and images, graphs, or charts.

The genAI revolution in legal practice is still in its early stages, and additional empirical research will be important to test, validate, and refine the results reported here across a wider range of matters, document types, workflows, and risk profiles. In particular, future studies should examine how tool performance varies with different prompts and review protocols, how well results generalize beyond a single closed matter, and what quality-control measures best support defensible use under prevailing discovery standards. Even with those open questions, the proof-of-concept findings in this Article provide a meaningful starting point: they suggest that, when appropriately deployed and monitored, genAI can exceed what is possible with manual review and search terms. In some respects, genAI is comparable to or exceeds the efficacy of TAR. In short, genAI may enhance the speed and consistency of large-scale review while preserving (and in some use cases improving) accuracy.

Accordingly, these results do not mark the end of the inquiry; rather, they mark the beginning of a conversation about where genAI belongs in the modern review workflow, what safeguards it requires (if any), and how lawyers can integrate these tools in ways that are aligned with core discovery obligations. Ultimately, genAI's value in eDiscovery will turn less on novelty and more on proof of its ability to be implemented in ways that are proportional, reliable, and fair. The challenge—and the opportunity—is to move beyond hype or hesitation and instead deploy genAI grounded in evidence and defensibility.