CASE STUDY

Optimizing Document Review Using Generative AI and Technology-Assisted Review

How AI-Augmented Review Strategies Delivered 90% Precision and 95% Recall

By Matthew Sinner, Esq., Bernie Gabin, and Young Yu

HAYSTACK®

Introduction

In eDiscovery, the task of reviewing vast quantities of electronic documents is one that requires a delicate balance between **speed**, **accuracy**, and **cost-efficiency**. Traditional manual review methods are time-consuming and costly, especially when the dataset spans hundreds of thousands, if not millions, of documents. The advent of advanced technologies like **Technology-Assisted Review (TAR)**, **Active Learning**, and, most recently, **generative AI (GenAI)** has given legal teams the tools to greatly increase the efficiency of document review, although they have also introduced novel challenges.

This case study explores the successful implementation of a hybrid workflow that combined **Relativity® aiR for Review™ (a Generative AI-powered document review solution), TAR**, and **conceptual analytics tools** to enhance document review for a complex legal matter. The case in question presented significant challenges, including a large document set, low overall richness, and a notable portion of documents classified as borderline relevant by aiR, requiring manual assessment. By strategically combining these technologies, the project, conducted by HaystackID®, achieved **major cost savings** and **enhanced review accuracy**.

The Challenges

The project encountered several key challenges in its early stages, which necessitated a thoughtful approach to workflow design.

1. **Limited Bandwidth for TAR Training:** At the outset of this project, it was anticipated that review would be conducted by a single attorney. A Prioritized Review workflow using Active Learning was considered, but the team concluded that too much review would be required for one user to properly train a TAR model in a timely fashion.
2. **Cost of Full GenAI Review:** Running an entire document set through AI analysis can be prohibitively expensive for some clients, especially when dealing with hundreds of thousands of documents. The initial goal was to find a more efficient way to use GenAI without having to process the full set.
3. **Challenging Random Sampling for Prompt Language Testing:** A key component to the use of any GenAI tool is the proper engineering of the prompt language, which instructs the tool on how to identify relevant content. This language is tested and refined on random samples of documents pulled from the document set. However, these samples may not have good conceptual coverage of the relevant issues or even contain enough relevant documents to properly evaluate the results. This problem is especially pronounced in low-richness situations, which can make it difficult to find examples of relevant documents.
4. **Managing Borderline Documents:** In GenAI document review projects, Borderline documents—those for which the AI tool cannot confidently make a relevance determination—are a frequent challenge. Handling these documents without incurring excessive costs from manual review is always a delicate balance. These documents were a significant issue in this case, as they required human intervention to resolve and would be expensive to address manually.

The Workflow: Using TAR to Augment GenAI

To address these challenges, the team designed a hybrid workflow that leveraged both GenAI (Relativity aiR for Review) and TAR.

Identifying the Review Universe

The initial dataset consisted of **891,527 documents**, of which **748,612** were eligible for analysis through aiR for Review. The remaining **142,915 documents** were excluded due to file type or size restrictions. Documents were filtered out if the text was smaller than **0.2 KB**, larger than **300 KB**, or was corrupt or otherwise unreadable.

Initial Prompt Language and Testing

The first step involved collaborating with outside counsel to draft the initial prompt language, which included case-specific background, key terms, relevant organizations, and descriptions of the types of documents the team was looking for. With this prompt in place, the team generated a test set of **384** randomly selected documents from the GenAI-eligible set, based on a **95% Confidence Level** and **+/- 5% Margin of Error**. This set was run through aiR and reviewed in full by counsel.

Counsel's review found only **17 relevant documents**, indicating just **4% richness** in the overall set. Using counsel's review as ground truth, all False Positives (documents coded Not Relevant by counsel but classified as Relevant by aiR), False Negatives (documents coded Relevant by counsel but classified as Not Relevant by aiR), and Borderline documents were sent to counsel for additional review. Using the Rationale and Considerations provided by aiR, this analysis provided additional context and insight to update the prompt language and improve subsequent runs.

Training the TAR Model

Counsel's Relevance coding of the first sample also provided the initial training for a TAR model created within Relativity's Review Center. This TAR model ranked the documents by their likelihood of relevance based on their similarity to documents coded as Relevant or Not Relevant. These scores were subsequently used to help target additional documents to **improve sample richness** as well as **define the scope of submissions to aiR**.

Second GenAI Sample

With the TAR model providing ranked predictions, the team created a second test sample that included the original **384 documents** plus **500 additional documents** that had received high relevance scores from the TAR model. The objective was to test the updated prompt language and to expand the pool of Relevant documents to improve the conceptual coverage of relevant content.

The second round was partially successful: the revised prompt identified more Relevant documents, but they were often very similar to each other, lacking true conceptual diversity. Additionally, many documents still fell into the Borderline category, meaning the AI could not definitively classify them. As before, these documents were sent to counsel for further review and analysis, which was then used to both refine the prompt language and retrain the TAR model.

Third GenAI Sample

In the third round, the team aimed to test the updated prompt language against a more diverse set of documents to identify a broader range of relevant content as well as compare its performance on documents that it had failed to properly classify in previous samples. Counsel provided exemplar documents from outside the review set, and the team used Relativity's conceptual analytics tools to find documents within the review set that were similar to these examples. The third sample contained a total of **500 documents**: **126** that had been classified as Borderline in previous runs, **174 documents** that counsel had coded as Not Relevant in previous runs, **52 documents** that counsel had coded as Relevant in previous runs, and **148** were new documents that were both conceptually similar to the exemplar documents and highly scored by the TAR model.

This round saw improved results, both in terms of providing definitive classifications and accuracy. Of the **126 documents** that had been Borderline with previous prompts, **28** now had classifications (a **22.22%** reduction). And of all non-Borderline documents, **90%** had classifications that accurately matched counsel's coding.

Finalizing the Prompt Language and Scaling Up

Following the three test samples, the team was confident the prompt language was performing well enough to submit a larger set. Rather than submit the entire review universe, the TAR model was leveraged to select additional documents for GenAI analysis. This would allow the team to **target documents** that had the highest likelihood of being Relevant, while avoiding submission of likely non-Relevant low-value records.

Documents were submitted to aiR in tranches of **2,000-4,000 records** at a time. These tranches included a mix of documents with high scores and a random selection of mid- and low-scoring documents for index health and conceptual coverage. Following each run, new aiR classifications were used as additional training for the TAR model.

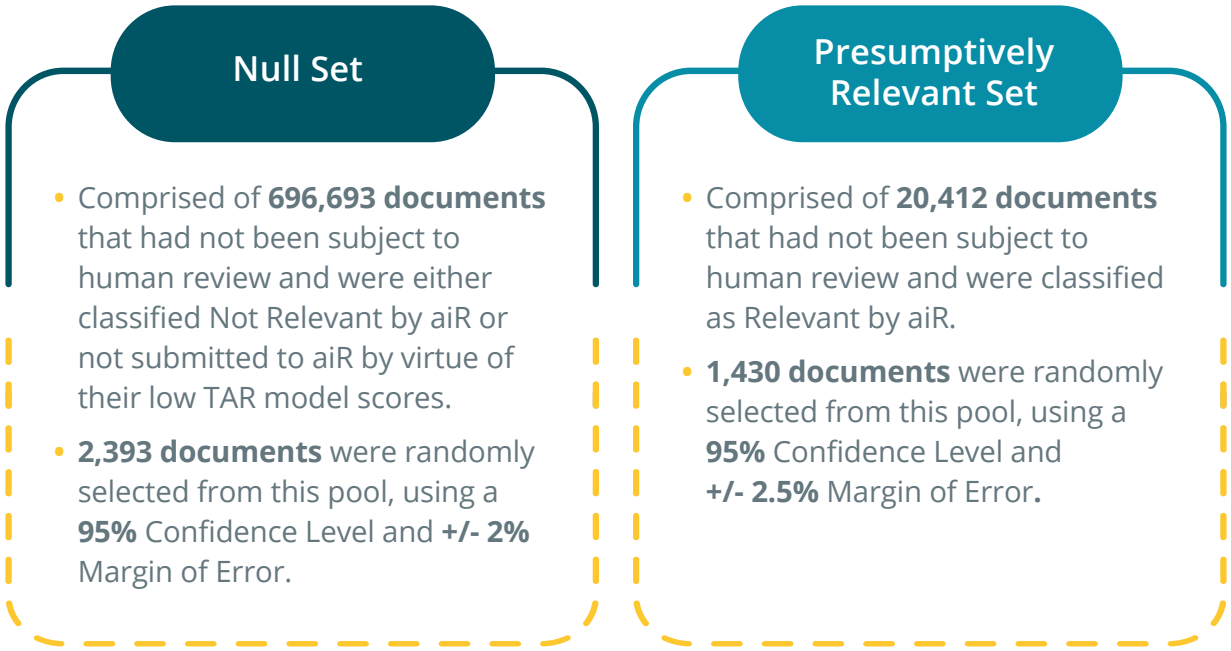
This training, along with further QC conducted by counsel, continuously improved the TAR model's accuracy and stability. When the model's development plateaued, a **cutoff score was selected**, and the remaining documents above that score were submitted to aiR for analysis.

Documents Sent for Manual Review

A contract review team was brought in to conduct a linear relevance review of the documents that could not be effectively analyzed through aiR for Review. This included documents excluded at the outset of the project as well as those classified as Borderline or Error by aiR. Review of the excluded documents was limited to individual records that had hit on search terms. Borderline and Error documents that were scored below the TAR cutoff were also excluded from manual review. In total, manual review was needed for **12,923 documents** within the GenAI-eligible review set.

Validation

When GenAI analysis and manual review had been completed, the team validated the results by analyzing two sets of documents:



Only four Relevant documents were found in the Null Set Sample, yielding an Elusion rate of **0.17%** (**1,165 documents**).

Within the Presumptively Relevant Set Sample, the validation confirmed that **86.29%** were truly Relevant. The **17,614 Relevant documents** from this pool, along with the **7,564 documents** that were coded Relevant through human review earlier in the project, resulted in a production set that contained **27,976 documents** with **25,178** projected to be truly Relevant, or **90.00% Precision** overall.

Recall was calculated from these metrics by dividing the total number of Relevant documents in the production set (**25,178**) by the total number of Relevant documents in the review universe (**26,343**). This resulted in a **Recall rate of 95.58%**, demonstrating that the AI and TAR workflow was both highly efficient and accurate in identifying Relevant documents.

Cost Savings

By narrowing the focus to the most Relevant documents based on TAR model scores, the team significantly reduced the volume of documents that needed to be run through Relativity aiR or human manual review. The full GenAI-eligible set of **748,612 documents** was reduced to just **54,380 documents** for analysis, cutting the GenAI processing volume by **over 90%**. This strategic reduction resulted in substantial financial savings. Importantly, this was achieved without compromising the quality, accuracy, or defensibility of the review process.

Conclusion

This hybrid approach, **combining GenAI with TAR**, delivered impressive results for the client. The workflow not only reduced the overall cost of the review but also improved its accuracy and efficiency.

While both tools perform the same core function—classifying documents—this process demonstrates that **they can be used simultaneously** in a way that is complementary rather than redundant. GenAI's ability to interpret and contextualize content is supplemented by TAR's ability to rank and prioritize documents.

There are still aspects of GenAI review projects that are likely to remain pain points. Good prompt revision and iteration requires subject-matter expertise and thoughtful analysis beyond simply coding documents. Many documents will not be eligible for GenAI analysis due to their file type or size. **Human review** will therefore continue to be a necessary element of eDiscovery projects.

Ultimately, this case study underscores that while GenAI holds great promise for transforming eDiscovery, its full potential is best realized when integrated with proven analytics tools, domain expertise, and a disciplined, iterative workflow. The path forward lies not in replacing human judgment or existing technologies, but in enhancing them through **smart, cost-effective augmentation**.

Learn More Today.

Contact us today to learn how HaystackID can help you solve your complex data challenges related to legal, compliance, regulatory, and cyber events.

About HaystackID®

HaystackID solves complex data challenges related to legal, compliance, regulatory, and cyber requirements. Core offerings include Global Advisory, Cybersecurity, Core Intelligence AI™, and ReviewRight® Global Managed Review, supported by its unified CoreFlex™ service interface. Recognized globally by industry leaders, including Chambers, Gartner, IDC, and Legaltech News, HaystackID helps corporations and legal practices manage data gravity, where information demands action, and workflow gravity, where critical requirements demand coordinated expertise, delivering innovative solutions with a continual focus on security, privacy, and integrity. Learn more at HaystackID.com.

About the Authors

Matthew Sinner, Esq.

Senior Generative AI And Analytics Consultant, HaystackID

Matthew Sinner joined HaystackID in 2020 and is currently a Senior Generative AI And Analytics Consultant. In this role, he develops and implements workflows utilizing structured analytics, conceptual analytics, machine learning, and GenAI. He works closely with project managers, clients, and review teams, providing guidance on appropriate strategies to address varying situations and ensuring the delivery of timely and effective solutions for client requests. Prior to joining HaystackID, Matthew was a Senior Analytics Consultant at NightOwl Global. He has been in various roles at NightOwl since 2015 and has been active in the eDiscovery space since 2012. Prior to joining NightOwl, Matthew worked with a political advocacy group and organized the intellectual property operations of a specialized construction materials company.

Bernie Gabin

Chief Data Scientist, HaystackID

In 2022, Dr. Bernie Gabin joined HaystackID and is currently the Chief Data Scientist on the Data Science team. In this role, he works closely with the company's CDS, John Brewer, to apply data-driven metrics to improve our procedures and develop custom AI/ML-empowered solutions for our clients. Prior to joining HaystackID, Bernie received his Ph.D. in physics from Brandeis University. His doctoral work in brain-computer interface systems and machine learning/artificial intelligence led him to work on AI/ML-focused projects for the US Patent Office, Northrup Grumman, and the National Security Agency. At HaystackID, he brings his expertise in signal processing, AI design, and data modeling to create novel data-driven solutions to our most challenging problems.

Young Yu

Vice President of Advanced Analytics and Strategic Solutions, HaystackID

Young Yu joined HaystackID in 2018 as a director and is currently the Vice President of Advanced Analytics and Strategic Solutions. In this role, Young is the primary strategic and operational adviser to HaystackID clients, focusing on the planning, execution, and management of eDiscovery activities. Young brings extensive experience to his position, having previously worked at IPRO Tech as a Professional Services Consultant/ Product Manager for Analytics, Wilmer Cutler Pickering Hale & Dorr LLP as a Team Lead/Litigation Support Coordinator, Chadbourne & Parke LLP as a Project Manager, and Ikon Office Solutions in various roles, including Data Engineer and Global Database Administrator. He holds certifications as a Brainspace Certified Admin, Analyst, and Specialist and is affiliated with Agile, Scrum, and Six Sigma. Young is proficient in a wide range of software and platforms, including Microsoft Office 365, Salesforce, SQL, Relativity, Ringtail, and many others.

Assisted by GAI and LLM technologies.

Source: HaystackID