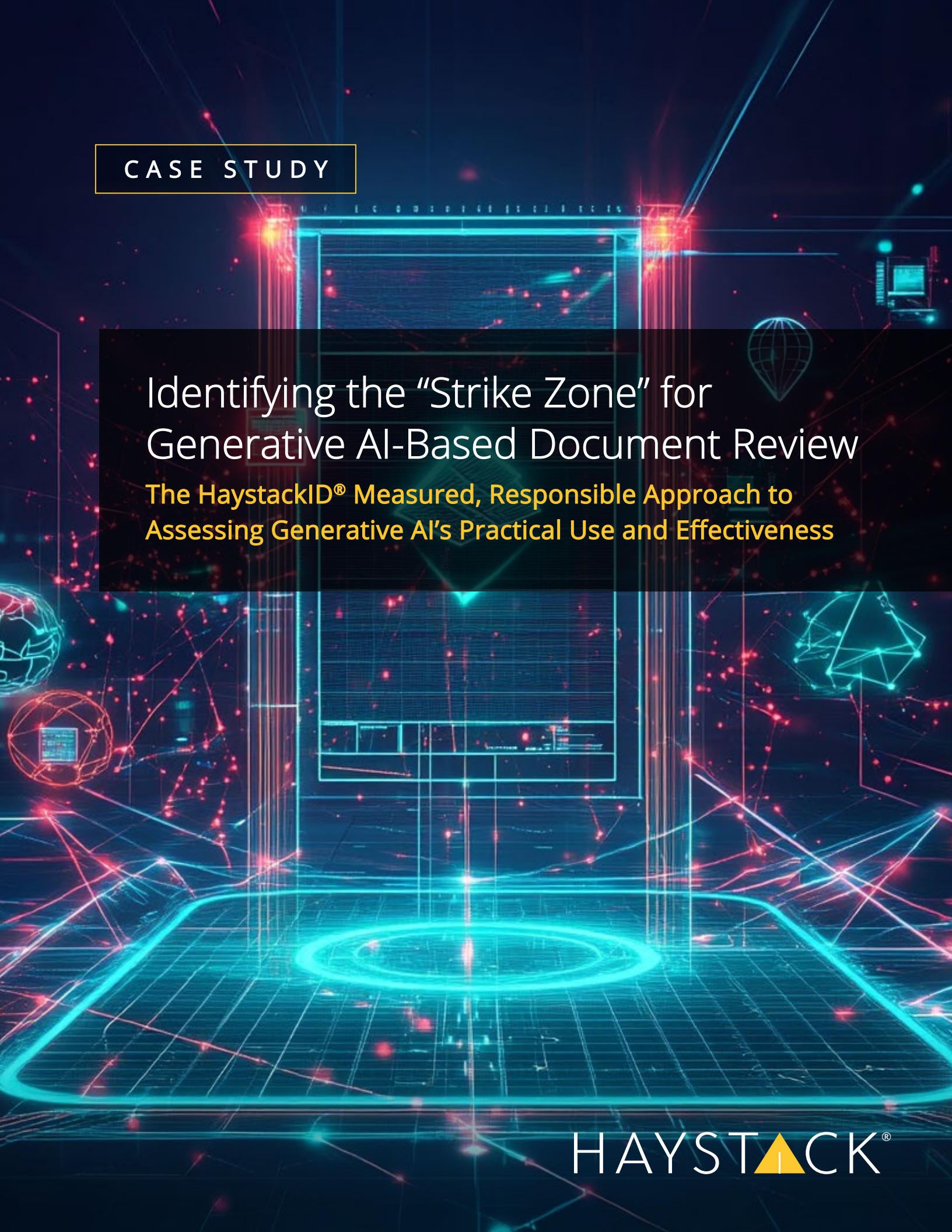

CASE STUDY

Identifying the "Strike Zone" for Generative AI-Based Document Review

The HaystackID® Measured, Responsible Approach to Assessing Generative AI's Practical Use and Effectiveness

HAYSTACK®

Introduction

Since **ChatGPT** exploded into the public consciousness in late 2022, the legal field has speculated on the possibility of using **Large Language Models (LLMs)** to execute first-level review of documents. The apparent excellent capability of LLMs to perform complex tasks on written documents seemed a natural match to the enormous and repetitive task of reviewing large numbers of documents for specific information. Although the capabilities of commercially available models in 2022 would have greatly limited the applicability of this exciting new technology, by the summer of 2024, LLMs had progressed to the point where **document review may be a practical application** of these programs.

Specializing in solving complex data challenges related to legal, compliance, regulatory, and cyber events, **HaystackID** has been committed to exploring the possibilities of these models in a measured and responsible way. Both in doing lab work to better understand the underlying technology and in evaluating products developed in the field, we have sought to understand the strengths and limitations of **Generative AI (GenAI)** as a tool for document review. In this paper, we'll be reviewing a specific case study using data from a specific matter that was previously conventionally reviewed that we re-reviewed using the relevance review available from **eDiscovery AI**, a legal technology company with tools for enabling GenAI document review.

Using the raw data in a low-pressure environment, without serious timeline pressure, we were able to repeatedly test and experiment with both the technology's uses and workflows to get the best results from the LLM review. The intent of this paper is to share our findings with the eDiscovery community to help inform them of the **strengths and limitations of using GenAI in review**.

Experiments

Overview

In order to test the effectiveness and potential of AI-based document review systems, we partnered with eDiscovery AI. They allowed us to use their system as an exemplar of the capabilities one can expect from such systems. While there are many different ways to build such systems, the one provided by eDiscovery AI should prove to be a reliable benchmark.

The majority of AI review systems are trained and tested using publicly available data sets, such as the Jeb Bush or Enron email corpora. However, when attempting to ascertain the value of these systems, it is imperative to test the system using real-world document review cases. To that end, we also partnered with a **U.S.-based securities firm**, who graciously allowed us to reprocess a case we had previously for them using conventional means through eDiscovery AI's system.

The Data Set

Our experimental data set consisted of a representational sample of **~30k documents** taken from the full case corpus of **~120k documents**. The 30k document subset was selected to be statistically similar to the full set and was employed in order to keep costs and run times manageable over repeated experimental runs.

As this document set had already been processed using conventional means, we were able to use the original review results as our **"ground truth"** to compare the AI system's answers. It is important to note that while the human review was treated as "truth" for the purposes of this analysis, it is understood that the original review also carried some amount of error. This conventional error was disregarded in this experiment as the focus was on comparing the AI system's performance to the conventional performance. However, a more in-depth statistical analysis that factored in both error rates would prove an enlightening direction for future work.

The original review included **nine separate requests**. While our experiment was focused on quantifying the AI system's overall ability to determine if a document was responsive or not, we did track the system's performance for each individual request as well. This allowed us to conduct further analysis of the system's topic-tagging capability.

The Document "Strike Zone"

As we began setting up the experiment, it became clear that we would need to modify our data set to conform to certain AI system limitations.

Due to the finite capacity of the systems context window, we were required to remove files above a certain size from the data set. The system was capable of processing **~48 kilo-tokens** at a time, and the context window needed to contain the document, request prompts, and system response. Estimating an average of **three (3) bytes per token** and reserving **1/3 of the window** for the prompt and results, we limited the maximum testable document size to **96 kilobytes**.

In addition, files that were too small and failed to provide the AI with enough context to properly process them were also disqualified. Any document that contained less than a sentence of words or was made up of purely numerical information was also removed from the data set.

Finally, the AI had difficulty dealing with overly structured or non-natural language files that lacked context (such as logs or automated system messages). Although the system proved to be very creative at deriving insights from documents with little context, it struggled when all context was lacking, such as CSV files with no headers. Thus, these documents were excluded as well.

Together, these limitations place bounds on document size (both upper and lower) and level of structure to create a **"strike zone"** for files that the AI system can optimally handle. Once all the documents that fell outside the strike zone were removed, our data set was reduced to **~26k documents**.

Sample Richness

In addition to selecting our sample to match the distribution of document types found in the full data set, the sample was constructed to contain **10 document subsets of 3k**, each of which was randomly selected from the full corpus to have a set richness ranging from **0% to 90%** based on the tags applied in the conventional review. This gave the entire **26k sample set** an aggregate richness of responsive documents of **40.26%**.

Training Data Set

For our initial experiment, we required one more data set to be defined. Documents were randomly sampled from the **26k** and then distilled into a set of **500 documents**, defined as our training set, that contained representative examples of each of the nine individual requests. Due to this requirement, the overall richness of the 500-document training set was slightly higher than the general data set, at **59.54%**.

Experimental Methodology

Prompt Training and Initial Test

We started by entering the original human review rubric for all nine requests directly into the AI systems prompt and then running it across the **500-document training set**. This provided us with a **"minimal effort"** baseline to compare future modifications.

We then had one of the subject matter experts (SME) who wrote the original rubric review of the AI system results work with our prompt engineers to modify the prompts to better align with the needs of the AI system. The new prompts were then run against the training set a second time.

A third and final iteration of the prompts was generated through the same process. The third iteration combined and modified the best-performing prompts from the first two training passes. After being run on the training set a third time, this set was deemed to perform suitably. The third prompt set was then run against the entire **26k test set**.

Inter-Run Variability

After completing our initial experiment, it was decided that we needed to get a better understanding of the inherent inter-run variability in the AI system. This was necessary to allow us to determine what changes to the prompts caused actual improvements to the results and what could be accounted for by random fluctuations. To test this, we conducted five runs of the third iteration prompt set on the training set while keeping all other variables constant. The distribution of the results from these runs allowed us to determine the standard inter-run variation.

Richness Subgroup Testing

As stated above, the full sample set was constructed from **10 smaller 3k samples** that were selected to have target richness levels ranging from **0% to 90%**. After removing the problematic files and running the remaining 26k documents, we reorganized the results by initial richness subgroup. This allowed us to determine what, if any, effect richness level would have on the AI system's performance.

Current Quality of AI Document Review

Experimental Results

Initial Test Results

1

As stated above, we ran three passes on the 500-document training set. After that, we got an overall classification rate of **91.84% recall** with **86.82% precision** on the training data.

We then ran the finalized prompts on the 26k document test set. For this set, we got overall classification rates of **90.92% recall** with **69.86% precision**.

2

These results are highly encouraging, as they indicate that the system is capable of results that emulate or exceed traditional human review. They also point to the fact that the system is tuned with a slight bias to recall over precision. This is a feature that should be accounted for in any workflows that utilize this type of AI system.

Inter-Run Variability

The inter-run variability test returned that the overall **recall** standard deviation was **1.05%**, and the overall **precision** standard deviation was **0.69%** across the five runs on the training data set. As the AI system processes each document individually, there is no reason to believe that this systemic variation should change between data sets, and thus, these values are equally applicable to all runs.

These values indicate that, while the system does have some systemic stochastic variation, it is small enough that prompt modifications that cause changes in excess of **3% recall** can be considered statistically significant to a **confidence of ~99%**. This is very important in allowing us to determine how many prompt training iterations are needed and when the modifications no longer have a significant impact.

Richness Subgroup Testing

As noted above, the AI system processes each document independently. Therefore, the richness of responsive documents in a data set should not affect the systems performance on a per-document basis. In order to confirm this, we determined the recall and accuracy for each of the richness fixed subgroups of the full **26k test set**. For this analysis, accuracy was used as opposed to precision because it gave more level grounds for comparison between very high and very low richness samples by factoring in both true positive and true negative classifications.

Unfortunately, we were forced to disregard the 0% richness subgroup, as it caused anomalous results due to a small number of statistics. Across the other nine groups **(richness 10% to 90%)**, recall was measured to have a **mean of 90.57%** with a standard deviation of **0.59%**. Our hypothesis was, therefore, born out for system recall, as the spread of these recall values is easily accounted for by inter-run variance.

However, we did see an increase in system accuracy as the richness increased. Accuracy values increased from **74.49% at 10% richness to 89.26% at 90%**. We believe this can be accounted for by the system's slight bias towards classifying documents as responsive, which is borne out by the high observed recall rates.

Inter-model Improvements

While we were running our experiments, eDiscovery AI began the rollout of their newest model version. We decided to extend our tests to this new model so that we could get a feeling for what could be expected out of such generational improvements. While the full analysis of the new model is outside the scope of this paper, the preliminary findings proved insightful.

Across three test runs on our **500-document training set**, we saw a decrease in **recall of 6%** but an **increase in precision** by the same amount as the previous model. This shows that new AI models and systems have the potential for meaningful improvements but also highlights the need to recalibrate and revalidate them as they come out. We should not naively expect general improvements in all metrics. In addition, certain models will be better suited for different jobs, and it is therefore important to select the correct one to fit a given matter.

Creative vs. Hallucination

No discussion of the application of AI systems would be complete without addressing the issue of **AI hallucinations**. Unfortunately, there are approximately as many definitions of hallucinations as there are papers in the field. After extensive discussion of which definition to use, we concluded that for the purposes of these experiments, it was irrelevant. While some definitions of hallucinations might include incorrect classifications, as in this case, we feel that venturing into that taxonomical debate might overshadow the practical conclusions of our work with semantics.

The tasks we are focusing on for these experiments boil down to binary classifications of **"responsive"** vs. **"non-responsive,"** or **"apply tag"** vs. **"do not apply tag."** We are concerned with the overall success of the system and not the specifics of why a misclassification was made. Therefore, any hallucinations the AI might generate are already accounted for in the system's overall error rate.

Conclusion

The principal conclusion of our research effort was that **using LLMs for some parts of document review is, indeed, practical and effective**. Although simply feeding an entire matter's review set into an LLM-based solution will produce suboptimal results, relatively simple cropping of the data set to give the LLM the technology only the files it is most qualified to review, as well as adjusting prompts to better align the review protocol with the LLM's capabilities, can indeed produce very high-quality review in a short period of time.

The concept of a **"strike zone"** is keenly important to our analysis. Documents that are too small to contain enough references for the model to gain a useful symbolic context do not work well. Documents that are too large to fit in a single context window, including both their prompt and output, carry risks with certain kinds of review requests. Documents that are under-structured, such that it is unclear what the context of the document is, are poor candidates for AI review, as are documents that are simply large quantities of structured data, such as expansive Excel spreadsheets of almost entirely numerical data.

Future Research

Our research laid out a number of directions for us to pursue.

Our experiments reveal potential issues on very low richness data set (<1%), specifically excessive false positives. Several explanations for this effect might be postulated, from a background false positive rate that becomes dominant in low-richness sets to as-yet undeciphered impacts of prompt construction.

Evaluating multiple issue tags in the same prompt revealed a non-intuitive effect in which adding more issues to a given prompt increased the accuracy of all prompts. How large this effect is, and if a "sweet spot" exists for a number of issues in a prompt, will require further research and experimentation.

Proposed AI-Enabled eDiscovery Workflow

Our testing has shown that AI-powered systems are far from black box solutions for eDiscovery review, and they can **significantly enhance review speed, accuracy, and efficiency.**

The initial workflow we developed, GenAI Assisted Review, would go as follows:

Step 1: Select a training sample from the full corpus of documents, making sure the sample contains examples of documents responsive to all requests. The exact sample size is a function of the overall corpus size, but initial testing indicates 500 – 1,000 documents should suffice for most matters.

Step 2: Have subject matter experts (SMEs) write up a review rubric for the matter.

Step 3: Have qualified SMEs apply the rubric to review and tag the training set in order to establish a "ground truth."

Step 4: Iteratively run the training set through the AI system, using the review rubric as the basis for the AI prompts. Between each run, have the SMEs review the results and modify the rubric/prompts to improve results. Care should be taken to improve recall while not causing too much degradation to precision. Repeat until system results are satisfactory to the client or until modifications no longer cause statistically significant changes to system performance. Initial testing indicates that three to five iterations should be sufficient.

Step 5: Run the entire corpus through the AI system using the final revision of the rubric/prompts.

Step 6: As recall was set as priority, the results of the run on the full corpus should weed out the majority of non-responsive documents and provide initial tagging. This reduced corpus can then be fed into a traditional review process such as Linear Review, TAR, or Active Learning. The AI curation thus enriches the corpus and acts as a force multiplier for the traditional methods, drastically improving their accuracy while greatly reducing the time and cost they require.

Future Workflows and Applications

Many other potential workflows have emerged. We expect to see new solutions entering the market in the autumn of 2024, allowing for privilege review, received production review, and new methods of quality control on AI review. These different missions have different parameters and requirements that will require their own sets of testing and validation but should prove to be greatly beneficial to the industry.

Drive Precise and Efficient ESI Discovery

Legal professionals face critical challenges in cybersecurity, information governance, and discovery. With **HaystackID® Core Intelligence AI™**, you can combat these challenges and gain the intelligence needed to assess rapidly expanding data sources.

Developed in partnership with eDiscovery AI and built on their state-of-the-art GenAI models, the technology combines unmatched natural language comprehension with deep contextual analysis to significantly reduce human error and enhance review accuracy compared to traditional AI, such as TAR and even Linear Review. **GenAI enables Core Intelligence AI to automate tasks such as data classification and categorization, transforming eDiscovery workflows, cutting costs, and maximizing efficiency.**

Learn More. Today.

Contact us today for more information on how HaystackID can help solve 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 events. Core offerings include Global Advisory, Data Discovery Intelligence, HaystackID Core® Platform, and AI-enhanced Global Managed Review powered by its proprietary platform, ReviewRight®. Repeatedly recognized as one of the world's most trusted legal industry providers by prestigious publishers such as Chambers, Gartner, IDC, and Legaltech News, HaystackID implements innovative cyber discovery, enterprise solutions, and legal and compliance offerings to leading companies and legal practices around the world. HaystackID offers highly curated and customized offerings while prioritizing security, privacy, and integrity. For more information about how HaystackID can help solve unique legal enterprise needs, please visit HaystackID.com.