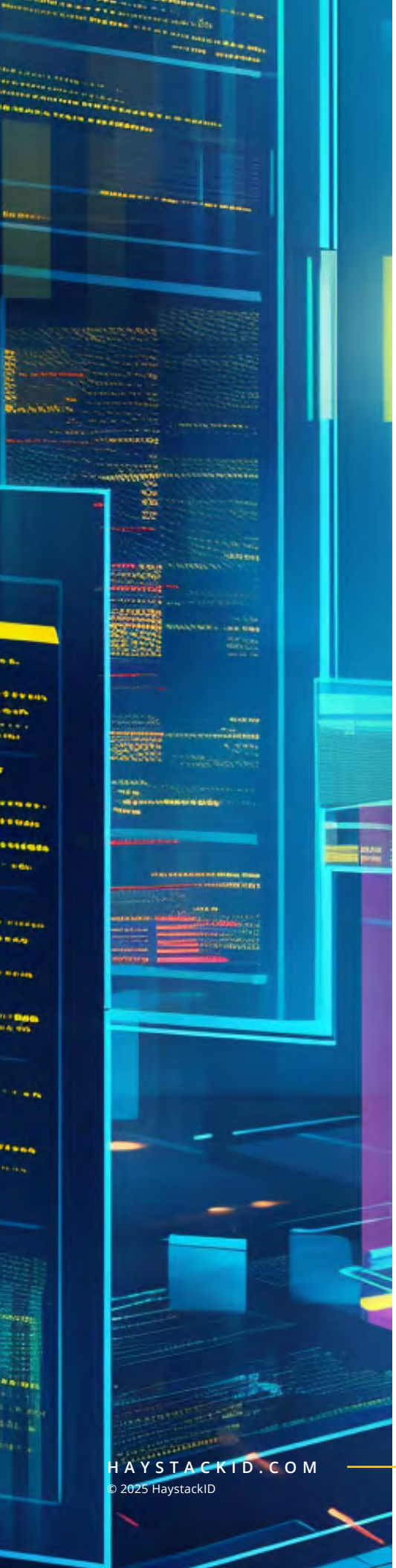
CASE STUDY

Tuning Generative AI-Based Systems to Enhance Document Review

A HaystackID® Analysis of aiR for Review

HAYSTACK®



Introduction

Since its emergence in late 2022, generative AI (GenAI) has been making waves in the legal industry, promising increased efficiency and accuracy in areas like document review. However, legal teams face a critical challenge: determining which AI tools deliver the most accurate, transparent, and defensible results. GenAI solutions have distinct architectures and methodologies, making quantifying risk and selecting the right tool for a given legal matter difficult.

To address this, HaystackID® conducted extensive, independent testing of Relativity aiR for Review, Relativity's GenAI-powered document review tool. We evaluated its real-world performance through structured benchmarking, iterative prompt engineering, and rigorous statistical analysis to assess its effectiveness and reliability. This hands-on, in-depth approach to AI validation is unique to HaystackID and reflects our commitment to providing defensible, high-performance AI solutions for legal professionals.

This paper focuses on three AI-driven workflows available in aiR for Review on the RelativityOne platform: **Issues Review**, **Relevance Review**, and **Relevance + Issues Review**. Our analysis compared these primary operating modes, revealing that while all three are effective, each is optimized for different applications, offering unique strengths depending on the legal team's specific review objectives.

Executive Summary

This paper aims to establish performance benchmarks for aiR for Review.
Our key findings include:

- All three tested operating modes within aiR for Review demonstrated strong performance. However, each mode is optimized to prioritize different evaluation metrics, illustrating the importance of selecting the appropriate mode based on a review's specific requirements.
- Effective prompt engineering is essential to optimizing recall, precision, and accuracy performance while minimizing the volume of documents requiring human review.
- Investing time in refining prompts at the outset of a project can significantly enhance AI efficiency by reducing the number of documents the system finds challenging to classify.

Evaluation of aiR for Review

aiR for Review leverages GenAI to replicate the decision-making process of a human reviewer. According to Relativity, the platform "uses generative AI to simulate the actions of a human reviewer, finding and describing relevant documents according to the review instructions you provide. It identifies the documents, describes why they are relevant using natural language, and demonstrates relevance using citations from the document."

To assess the system's capabilities, we analyzed its three distinct operating modes, or analysis types:

1. **Relevance Review:** This mode determines whether documents are relevant to a specific case or legal matter based on predefined criteria. It is beneficial for identifying documents responsive to a production request.
2. **Issues Review:** This mode is designed for more granular classification. It categorizes documents according to specific issues defined by the user, similar to a conventional review guide. For example, it can identify documents related to coercion, retaliation, or both.
3. **Relevance + Issues Review:** This mode combines the functionality of both Relevance Review and Issues Review, classifying documents for general relevance and specific issue categorization. While it might be expected to perform similarly to the individual modes it combines, our analysis found that it yields distinct results, suggesting unique weighting and processing mechanisms.

Each mode offers valuable functionality, but our findings indicate that selecting the appropriate analysis type is critical to achieving optimal results based on a given review's objectives.

Experimental Method

The Dataset

We utilized the same dataset for this evaluation to ensure consistency with our previous experiments. Our test set consisted of a representative sample of approximately ~30,000 documents drawn from a full corpus of ~120,000 documents that our team had previously reviewed using conventional methods. While we acknowledge that human-based review is inherently imperfect, assessing how these errors compared to those generated by AI was beyond this study's scope. As such, we treated the original review results as the ground truth against which we measured the AI system's performance.

We applied the same "Strike Zone" parameters to the dataset to maintain alignment with prior experiments. The Strike Zone defines a set of criteria ensuring that only documents suitable for AI-based analysis are included, improving the system's performance and comparability across experiments. This involved:

- Excluding documents too large to fit within the AI model's context window, as they exceeded the system's processing limits.
- Filtering out documents that were too short to provide meaningful context. This lack of context makes it difficult for the AI to assess relevance.
- Removing highly structured or non-natural language documents, such as logs, system-generated messages, or spreadsheets. Since Large Language Models (LLMs) are primarily trained on human language, they are less effective at analyzing these types of content, which are typically better suited for specialized human review pipelines.

After applying the Strike Zone criteria, we were left with a refined test dataset of approximately ~26,000 documents, ensuring the AI's performance was measured on the content it was best equipped to process. The original conventional review included nine distinct issue requests. While our primary evaluation focused on the AI system's ability to classify document responsiveness rather than identifying specific issues, aiR for Review Issues Review mode explicitly tags documents by issue. As a result, we tracked both its performance on individual issue categorization and its overall classification accuracy.

We extracted a smaller training set from the test dataset to develop effective prompts for the AI system. This subset included representative examples for each of the nine issue requests. We added approximately 150 non-responsive files to the training set, bringing the total number of documents in the set to 500.

Richness of the Dataset

The original 30,000-document sample included ten separate 3,000-document subsets, each randomly selected from the full 120,000-document corpus. These subsets were structured to reflect varying levels of document responsiveness, ranging from 0% to 90% richness, based on the tags applied in the conventional review process. After applying the Strike Zone criteria—removing documents that were too large, too small, or too highly structured—our team reduced the final test dataset to 26,000 documents. The overall proportion of responsive documents (or richness) within this refined dataset was 40.26%. For the 500-document training set, a higher richness level was necessary to ensure that all nine issue categories were adequately represented. As a result, the training set had a richness of 59.54%, providing the AI system with a well-balanced foundation for learning and classification while maintaining alignment with the broader test dataset.

The Procedure

We followed the standardized testing procedure developed during our earlier experiments while making minor adjustments to accommodate each aiR for Review operating mode.

Initial Experiment - Issues Review

While our primary objective was to evaluate the AI system's overall ability to determine document relevance, we first assessed the Issues Review mode in-depth due to its information-dense output. This prioritization allowed for a more granular analysis of how the system responded to specific issue descriptions and provided an opportunity to refine prompts iteratively.

Our approach followed a structured, multiphase process:

1. **Baseline Testing with Original Review Guide Prompts**
   To establish a starting point, we used prompts from the original human review guide and entered each of the nine issue requests as separate inputs into the AI system. These prompts were tested against the 500-document training set, generating initial baseline results.

2. **Refinement with AI-Optimized Prompts**
   Next, we introduced a revised set of prompts specifically crafted for AI-based review by one of the subject matter experts (SME) who authored the original rubric. We then ran these AI-optimized prompts against the training set for a second round of evaluation.
3. **Final Iteration with Performance-Tuned Prompts**
   The best-performing prompts were further refined based on insights from the first two test passes. This third and final iteration was tested against the 500-document training set to validate improvements in classification accuracy.
4. **Full-Scale Testing on the 26,000-Document Set**
   Once the optimized prompts were finalized, we applied them to the full 26,000-document test set to simulate a complete review pipeline. For reporting purposes, any document tagged with one or more issue requests was classified as relevant in the general case.

This iterative approach enabled us to systematically improve the AI system's performance while gaining deeper insight into how prompt engineering influences classification outcomes.

Inter-Run Variability

Following the initial experiment, we wanted to better understand the inherent inter-run variability within the AI system. This step was critical in distinguishing genuine improvements in results from variations caused by random fluctuations.

We conducted four additional runs using the third iteration of prompts on the 500-document training set to achieve this, ensuring that all other variables remained constant. **By comparing the results across these runs, we were able to:**

- **Quantify** the system's natural variability when processing the same dataset under identical conditions.
- **Determine** whether observed changes in performance were due to refinements in prompt engineering or random fluctuations.
- **Calculate** standard inter-run variation across the initial run and the four additional runs, allowing us to incorporate these values as error margins in our derived metrics.

This approach provided a more rigorous and statistically grounded evaluation of how prompt modifications influenced AI performance, ensuring that any reported improvements were meaningful and repeatable.

Relevance Review and Relevance + Issues Review

To ensure consistency in evaluation, we repeated the Prompt Engineering and Inter-Run Variability tests for both Relevance Review and Relevance + Issues Review modes.

- **In the Relevance Review runs**, we entered all prompt text for each version into the single Relevance field within the Relevance Review mode.
- **In the Relevance + Issues Review runs**, we placed each request's text in a separate Issue field and entered the general review guide overview into the Relevance field.

After collecting and analyzing these results, we determined that the training dataset runs provided sufficient data to compare the performance of the different modes. As a result, we deemed additional full-scale runs on the 26,000-document test set unnecessary, avoiding the associated time and cost.

Borderline Document Mitigation

A key feature of aiR for Review is its ability to flag documents it cannot confidently classify as Borderline. Additionally, some files may fail to process due to technical issues, which the system labels as Errors. Regardless of the reason, both categories require human intervention, making it essential to minimize their occurrence to maximize the efficiency of AI-assisted review.

During our Issues Review run on the full test set, the issue with the highest number of Borderline classifications was 4(j), with 3,350 flagged documents. To assess whether targeted prompt engineering could reduce the number of Borderline cases, we drew three random samples of 300 documents from the 4(j) Borderline dataset. We also had the SME who authored the original prompts refine the 4(j) prompt, using aiR for Review's explanations of why it was unable to classify these documents. From there, we created an updated "Request 4(j) Version 3.1" prompt and tested it on the three sample sets to evaluate its effectiveness in reducing the number of unclassified files.

This process allowed us to assess whether iterative refinement of AI prompts could improve classification accuracy and reduce the volume of documents requiring human review.

Experimental Results

Metrics Overview

To comprehensively assess the performance of the tested systems, we applied several standard metrics commonly used to evaluate binary classifiers. Since this experiment specifically focused on the system's ability to classify documents as Responsive or Non-Responsive, treating it as a binary classification problem was appropriate. It is important to note that documents flagged as Borderline or Errors—where the system could not make a definitive determination—were excluded from Recall, Precision, F-score, and Balanced Accuracy calculations. Instead, the percentage of these unclassified documents was reported separately, as they represent cases requiring additional human intervention.

Recall

Recall, sometimes called Sensitivity, is the ratio of True Positives returned by the system to All Positives in the dataset (i.e., True Positives plus False Negatives). In essence, it is the percentage of responsive documents the system found.

Precision

Precision is the ratio of True Positives returned to All Positives returned by the system (i.e., True Positives plus False Positives). This metric indicates the trustworthiness of the system's responsive call and assures the system does not simply return massive numbers of documents to ensure high Recall.

F-Score

The F-score is the harmonic mean of Precision and Recall. It allows for a concise way to compare the performance of different systems across both primary metrics.

Balanced Accuracy

Balanced Accuracy measures how well the system classified both Relevant and Non-Relevant files while being renormalized for imbalances in sets with low richness. Our team deemed this metric important to help better gauge how well the system performed in rejecting non-responsive files for such low richness issues.

Borderline + Error Percentage

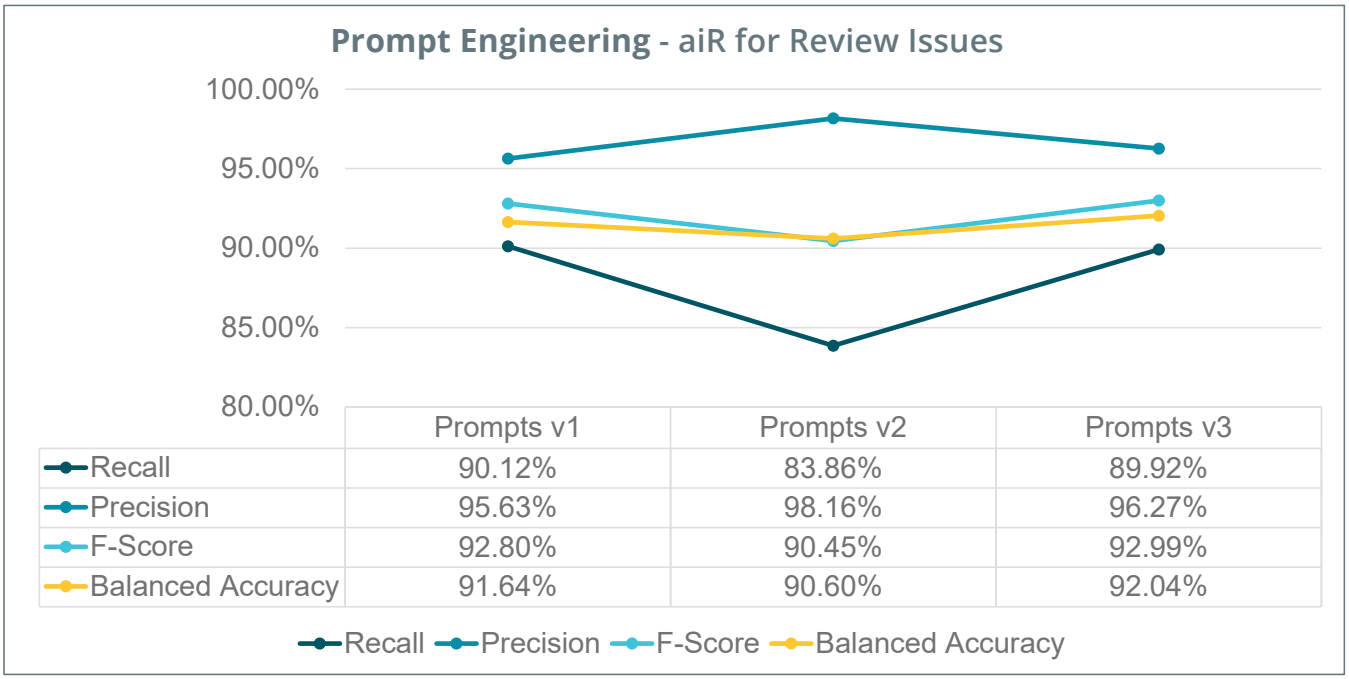
As observed above, aiR for Review would characterize any documents that could not be classified as Borderline or Error files. For each run, the number of such documents requiring further manual human review as a percentage of the total dataset was calculated.

Variance

Due to AI systems' inherently stochastic nature, we expected a certain difference in the results of identically initialized runs. It is essential to characterize this inter-run variance to understand if changes made to the initial conditions (i.e., prompt engineering) cause changes to the results or if any observed changes can be accounted for by chance.

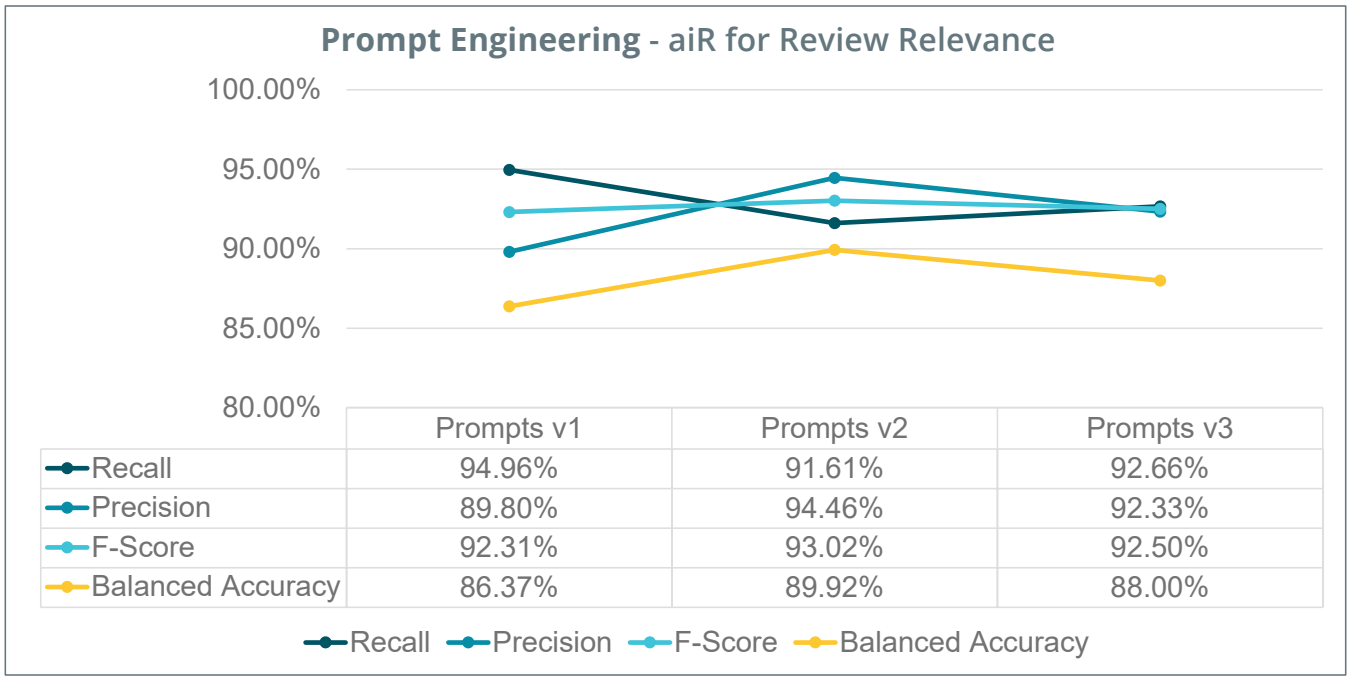
Prompt Engineering Runs

The results from the prompt engineering runs can be summarized in Figures 1 through 3 for each aiR for Review operating mode we tested. Figure 4 summarizes the Manual Review Percentage for all three modes across all three prompt versions. These Figures clearly illustrate the tradeoffs made during the prompt engineering process to optimize the balance between Recall and Precision while minimizing the number of files requiring further review.



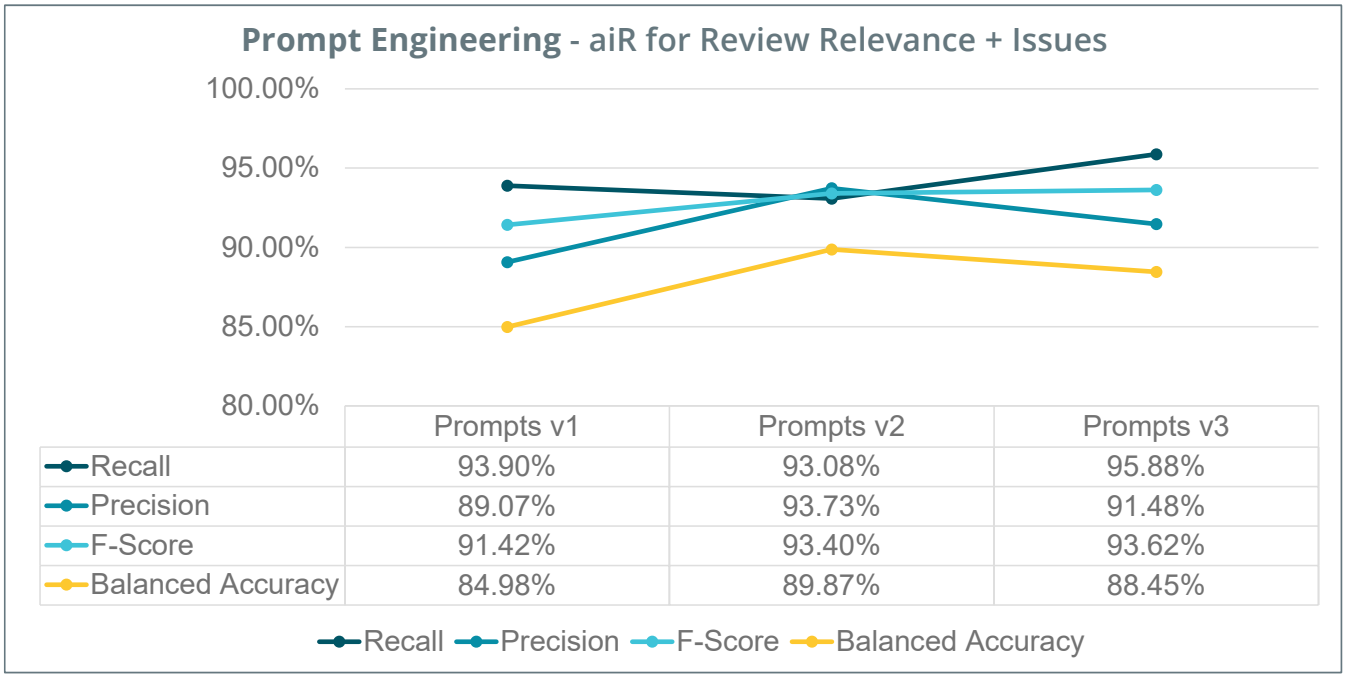
	Prompts v1	Prompts v2	Prompts v3
Recall	90.12%	83.86%	89.92%
Precision	95.63%	98.16%	96.27%
F-Score	92.80%	90.45%	92.99%
Balanced Accuracy	91.64%	90.60%	92.04%

Figure 1 - Summary of Prompt Engineering runs for aiR Issues mode



	Prompts v1	Prompts v2	Prompts v3
Recall	94.96%	91.61%	92.66%
Precision	89.80%	94.46%	92.33%
F-Score	92.31%	93.02%	92.50%
Balanced Accuracy	86.37%	89.92%	88.00%

Figure 2 - Summary of Prompt Engineering runs for aiR Relevance mode



	Prompts v1	Prompts v2	Prompts v3
Recall	93.90%	93.08%	95.88%
Precision	89.07%	93.73%	91.48%
F-Score	91.42%	93.40%	93.62%
Balanced Accuracy	84.98%	89.87%	88.45%

Figure 3 - Summary of Prompt Engineering runs for aiR Relevance + Issues mode

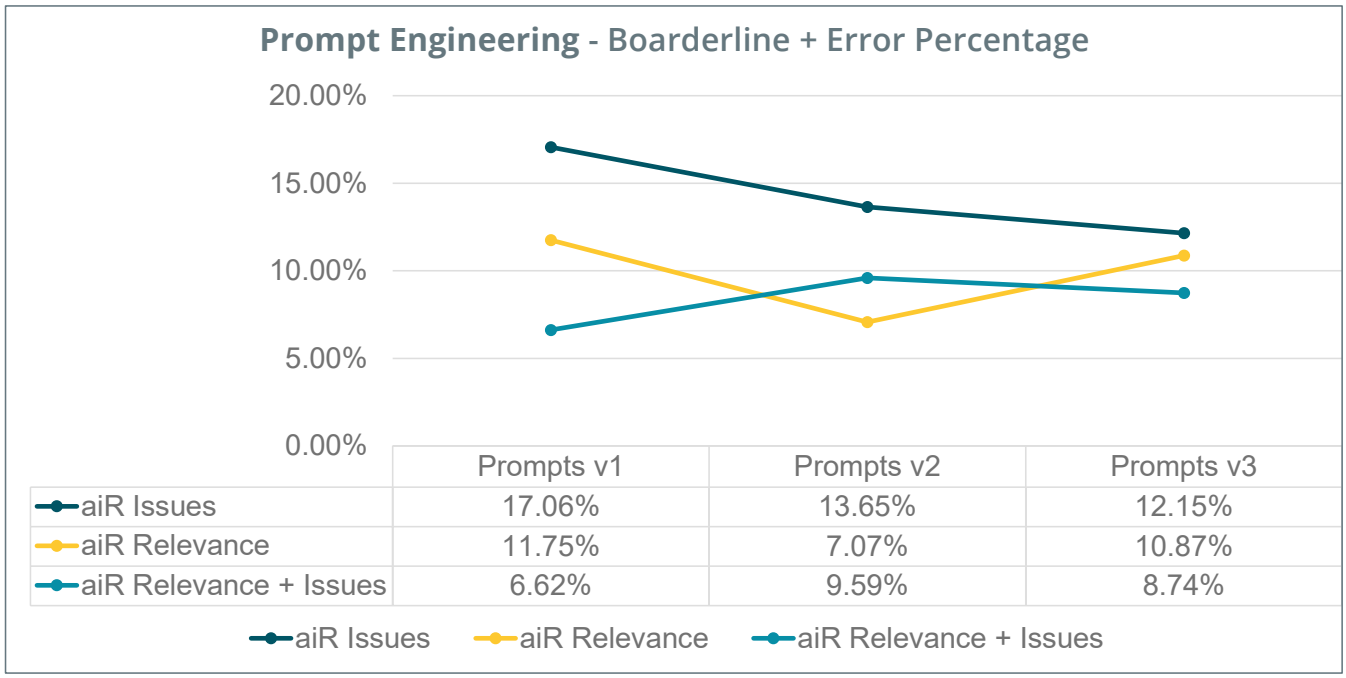


Figure 4 - Summary of Manual Review Percentages across all Prompt Engineering runs

Inter-run Variance Results and Full Test Set Run

Our goal was to place values on the system's inter-run variance. Therefore, we ran each finalized prompt set (version 3) five times on the training dataset and took standard deviations of the results. The values reported below, in Table 1, are the average of these five runs, and the errors provided are the standard deviation. Figure 5 summarizes these tests' results.

Table 1 and Figure 5 included the full results from the 26,000-document test set run for easy comparison. These results clearly show a drop in performance between the training and test runs that cannot simply be accounted for by run variance. However, this decrease does not push the results below a usable threshold. Future research will need to be undertaken to fully characterize the nature and amount of this decrease and best practices to mitigate it.

	Recall	Precision	F-Score	Balanced Accuracy	Manual Review
aiR Issues	89.57 +/- 0.63 %	97.53 +/- 0.44 %	93.38 +/- 0.38 %	92.45 +/- 0.49%	10.25 +/- 1.28 %
aiR Relevance	92.80 +/- 0.50 %	92.49 +/- 1.02 %	92.64 +/- 0.57%	88.43 +/- 1.01%	9.25 +/- 1.07%
aiR Relevance + Issues	94.76 +/- 1.03%	91.69 +/- 0.36%	93.20 +/- 0.54%	88.47 +/- 0.58%	7.29 +/- 1.18%
aiR Issues – 26k Test Set	87.77 +/- 0.63%	89.81 +/- 0.44 %	88.78 +/- 0.38 %	91.30 +/- 0.49%	13.95 +/- 1.28%

Table 1 - Summary of Inter-Run Variance Results and Full Test Set run

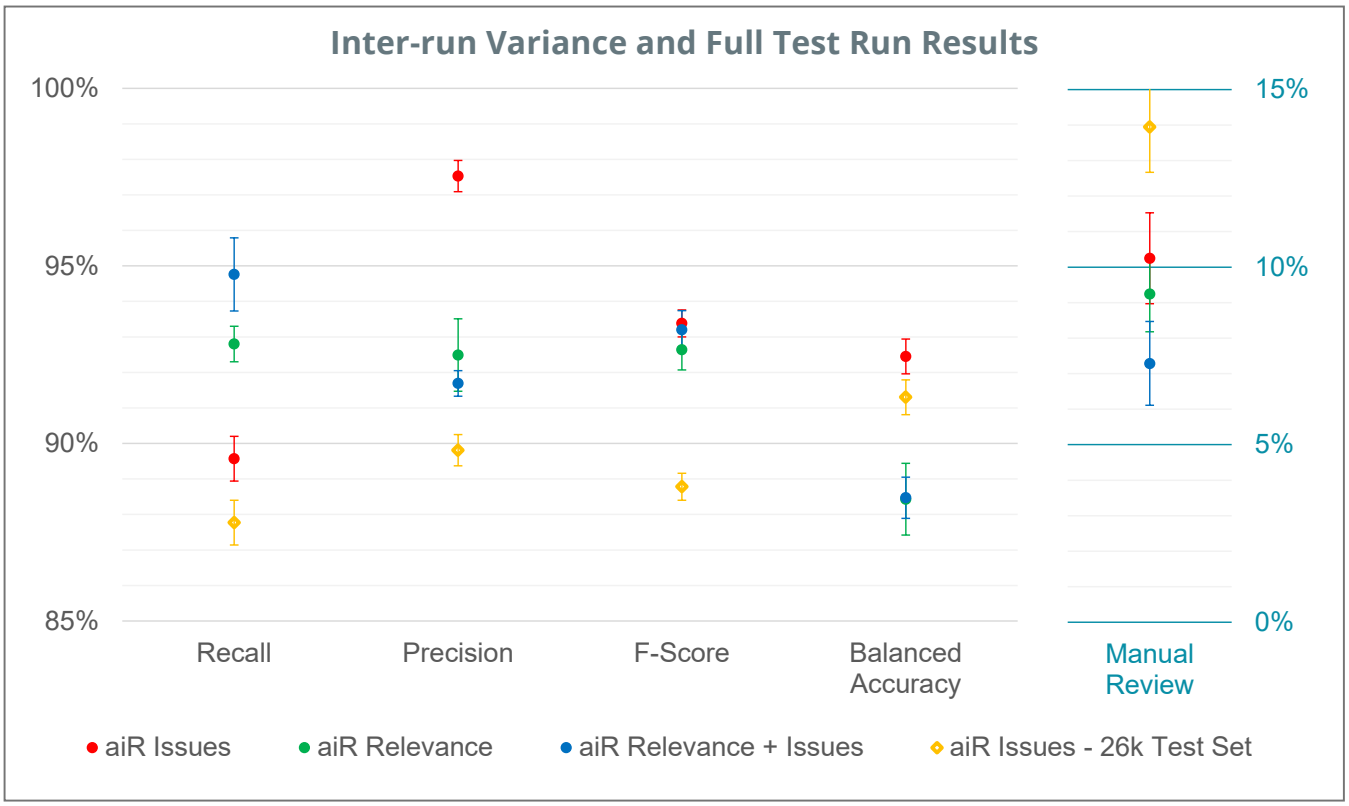


Figure 5 - Summary of results for the Inter-run Variance and Full Test Set experiments

Borderline File Mitigation Results

Figure 6 summarizes our Borderline File Mitigation test. As stated above, these results are for a new version of the prompt for Issue 4(j) applied to three different 300 document samples drawn from the 3,350 4(j) Borderline documents reported from the full 26,000-test set. Based on our inter-run variance metrics for the Issue Review operating mode, all experimental runs have an error of +/-1.28%.

The results clearly show that with even minimal prompt engineering, the total number of documents requiring further human intervention can be drastically reduced. However, they also indicate that the amount of that reduction is highly dependent on the specifics of the files being addressed.

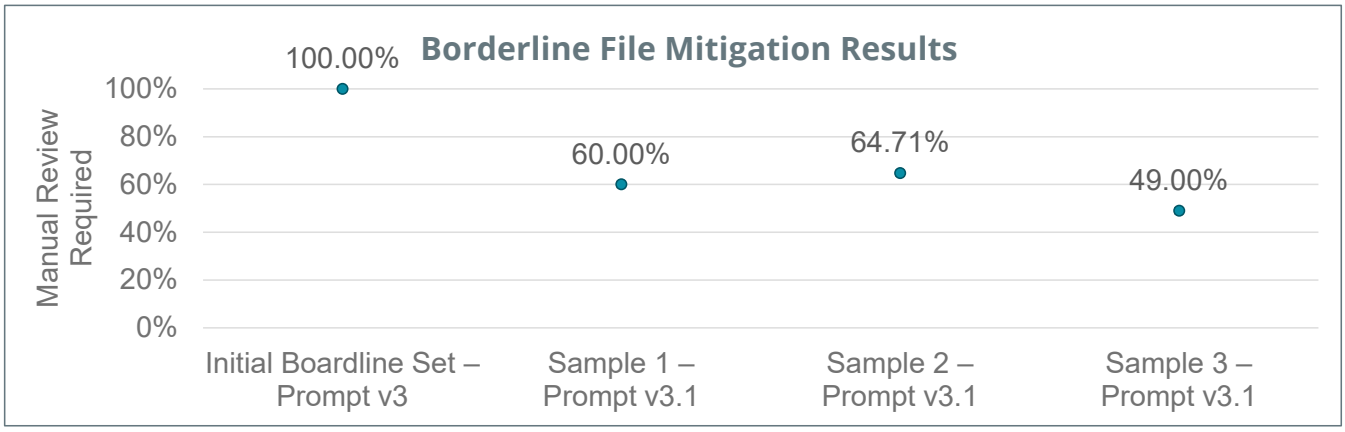


Figure 6 - Summary of the Borderline File Mitigation experiment.
Experimental results for samples 1, 2, and 3 have 1.28% error bars on them.

Conclusion

Our experiments demonstrate that aiR for Review offers significant flexibility, allowing users to tailor AI-assisted review to the most critical aspects of their specific matters. With multiple performance metrics to consider—such as Recall, Precision, and Borderline fraction—optimizing results requires a careful balance. While this study provides insights into how you can adjust these metrics for better outcomes, the ideal balance will vary from one review to another.

Like all GenAI-powered review tools, aiR for Review requires a robust testing phase to ensure reliable performance. Its intuitive user interface makes statistical review more accessible than many comparable solutions. However, deeper analysis beyond the standard interface metrics remains essential to achieving optimal results. Notably, while reducing the number of Borderline classifications is beneficial for minimizing human review, it may not always be feasible across diverse real-world document sets.

aiR for Review operating modes demonstrate the most value when used in combination. For example:

- **Issues Review** mode enhances precision, though sometimes at the expense of recall.
- **Relevance + Issues Review** mode significantly reduces the need for human review, at the expense of some accuracy and precision.

Additionally, while the Borderline classification feature is valuable, it heightens the importance of thorough prompt engineering. A moderate prompt error that increases the number of confidently misclassified documents can be costly, as such issues typically only surface in post-analysis sampling. While aiR for Review represents a highly advanced implementation of GenAI for document review, it still faces the same fundamental challenges as all AI-driven review tools—and document review in any form.

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.