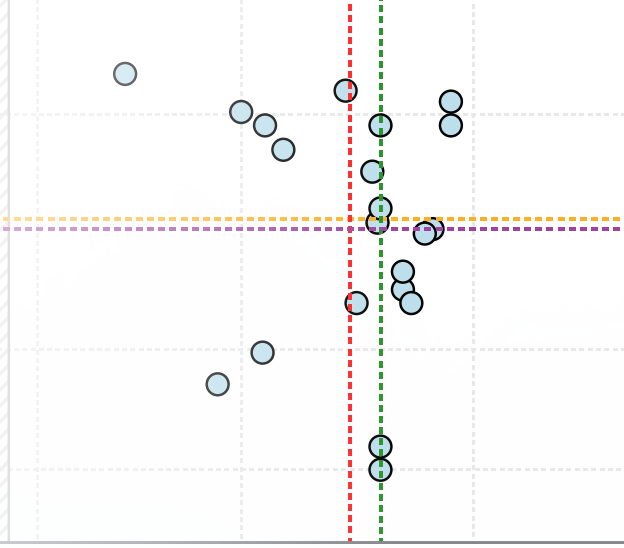
JND

BEYOND THE HYPE: A FIELD STUDY IN AI-BASED DOCUMENT REVIEW



INTRODUCTION

It can be difficult to parse fact from fiction around GenAI in the legal industry. “Does it work?” is a question with as many answers as there are use cases for GenAI, from research and drafting to discovery and trial support. Of all the use cases, document review might be the only one with an established framework for benchmarking its performance, thanks in large part to the groundwork laid by TAR.

To that end, you’ve probably seen academic studies, case studies and user testimonials demonstrating that GenAI-based review is capable of easily passing the 80% “recall test” often used to validate TAR projects. In fact, GenAI users are describing 90%+ recall as the “new normal”.

According to Chris Haley of Relativity, “Across hundreds of customer using aiR on thousands of projects, we’re consistently hearing reports of 90%+ recall, easily surpassing validation requirements typically relied on for TAR, and outperforming alternative review methods, with far less effort.”

Jim Sullivan , founder of eDiscovery AI, states “As someone who has used TAR for over ten years on hundreds of cases, without exception, GenAI is the best classifier available on the market. We consistently see users reaching 90%+ recall and precision on live matters.”

With a rapidly growing user base, I wondered if it might be possible to expand on the traditional case study, and take a more data-driven look at how GenAI review is performing in the field, across multiple users and projects. I reached out to law firms, corporations, service providers and software to survey if they were using GenAI for document review, and if they’d be willing to share the results.

Unfortunately, lawyers aren’t quick to share sensitive details about their discovery process with strangers on the internet. But, with a touch of Minnesota nice and several legally binding NDAs (cannot connect results to sources), I was able to gather validation data from twenty one review projects across four different software platforms.

THE DATA SOURCES

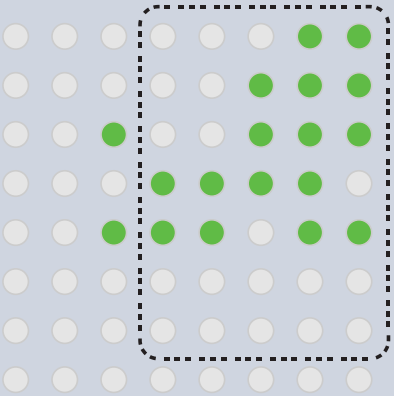
- Contributions came from corporations (4), law firms (9), legal service providers (4), and software vendors (4).
- Of the 21 projects, 16 involved real cases, while 5 were from proof-of-concept (POC) trials or publicly available studies.
- The review types include responsiveness (14), privilege (5), and personally identifiable information (PII) (2).
- The results come from four different GenAI-based review platforms.

REFRESH ON RECALL AND PRECISION

As a quick refresher, the most common way to evaluate TAR or GenAI-based review is by measuring **recall** and **precision**. **Recall** represents the percentage of responsive documents that were correctly identified as *responsive*. **Precision** measures the percentage of documents labeled as *responsive* that are actually responsive.

Put simply, **recall** answers the question, *"Did you find everything?"* while **precision** asks, *"How much extra junk did you pick up?"*

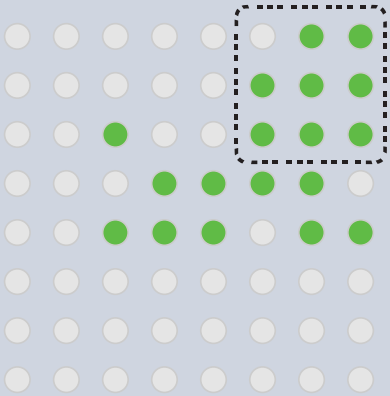
High Recall, Low Precision



Recall = 88.9%
(16 out of 18 responsive documents were found)

Precision = 45.7%
(16 out of 35 documents found were responsive)

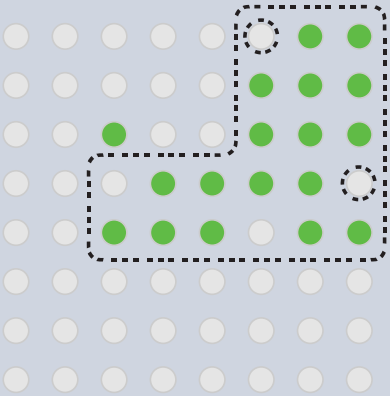
High Precision, Low Recall



Recall = 50%
(only 9 out of 18 responsive documents were found)

Precision = 88.9%
(8 out of 9 documents found were responsive)

High Recall, High Precision



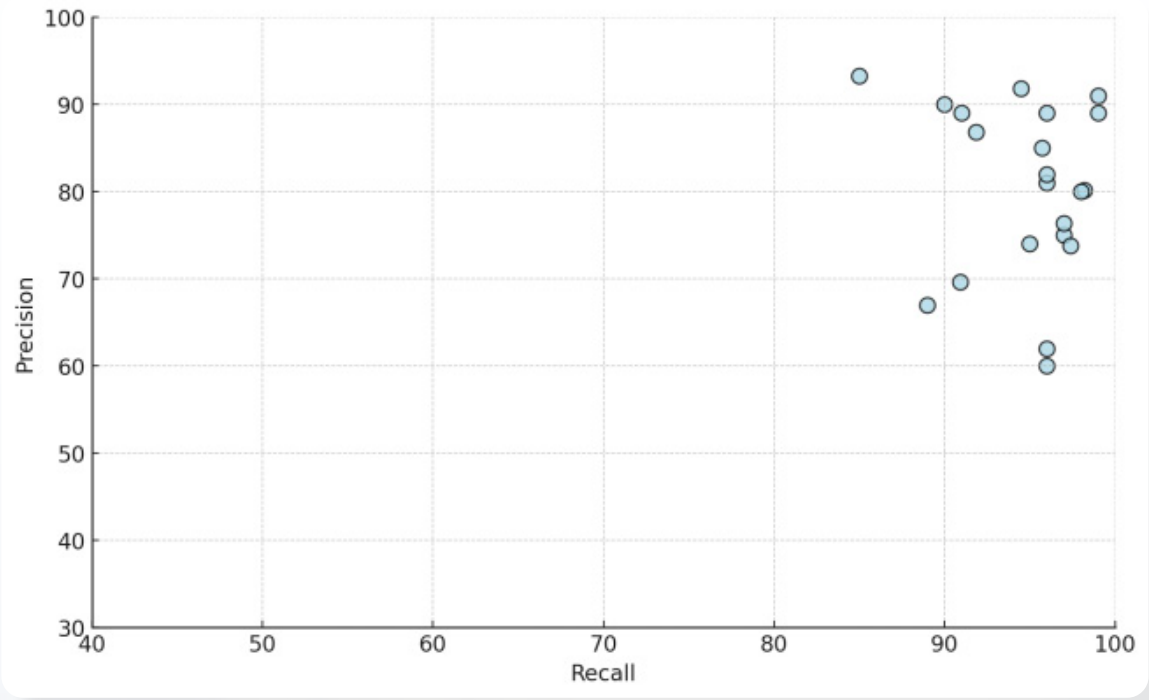
Recall = 94.4%
(17 out of 18 responsive documents were found)

Precision = 89.4%
(17 out of 19 documents found were responsive)

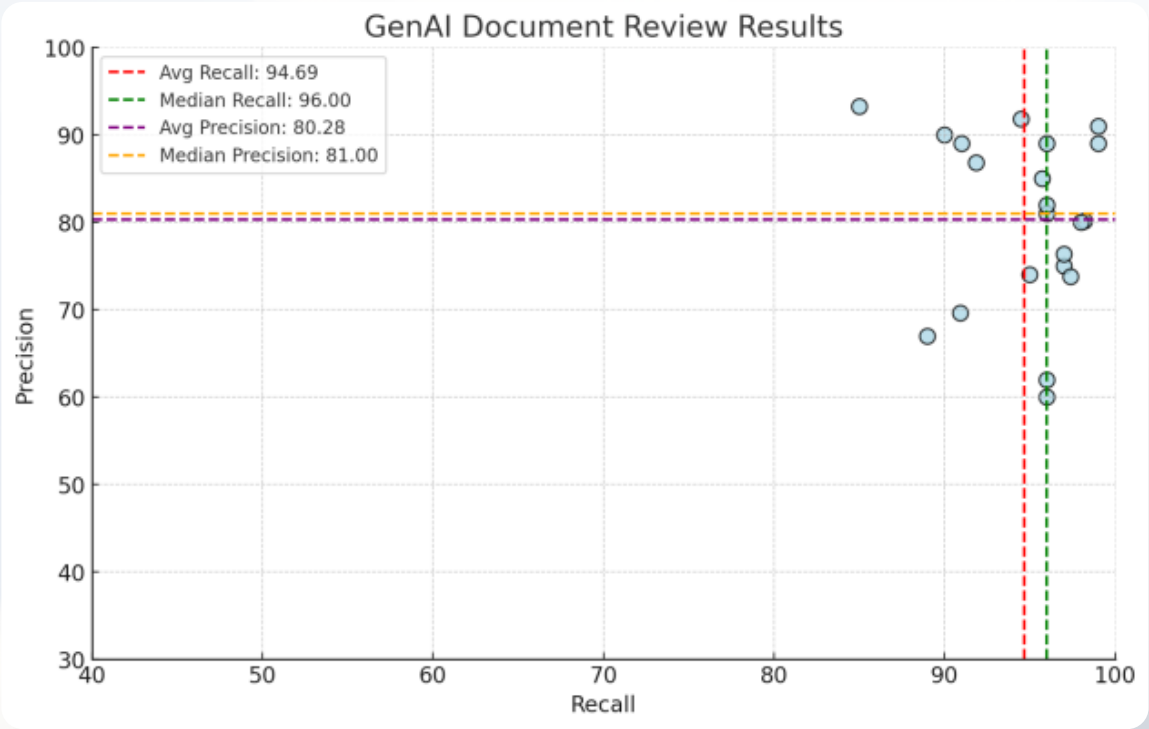
For simplicity, all data in this article is being presented as "point estimates" for recall and precision, as opposed to including margin or error (which wasn't always available).

THE RESULTS

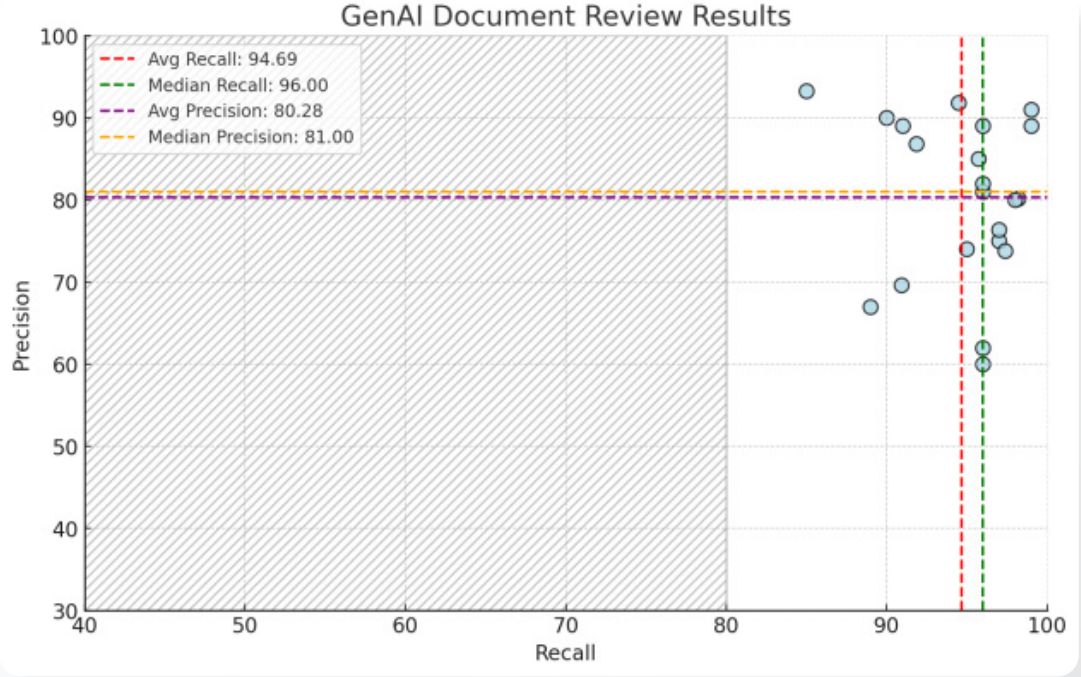
The data points are plotted below with **recall** as the **x-axis** and **precision** as the **y-axis**.



Here, I've added the average and median lines for recall and precision.



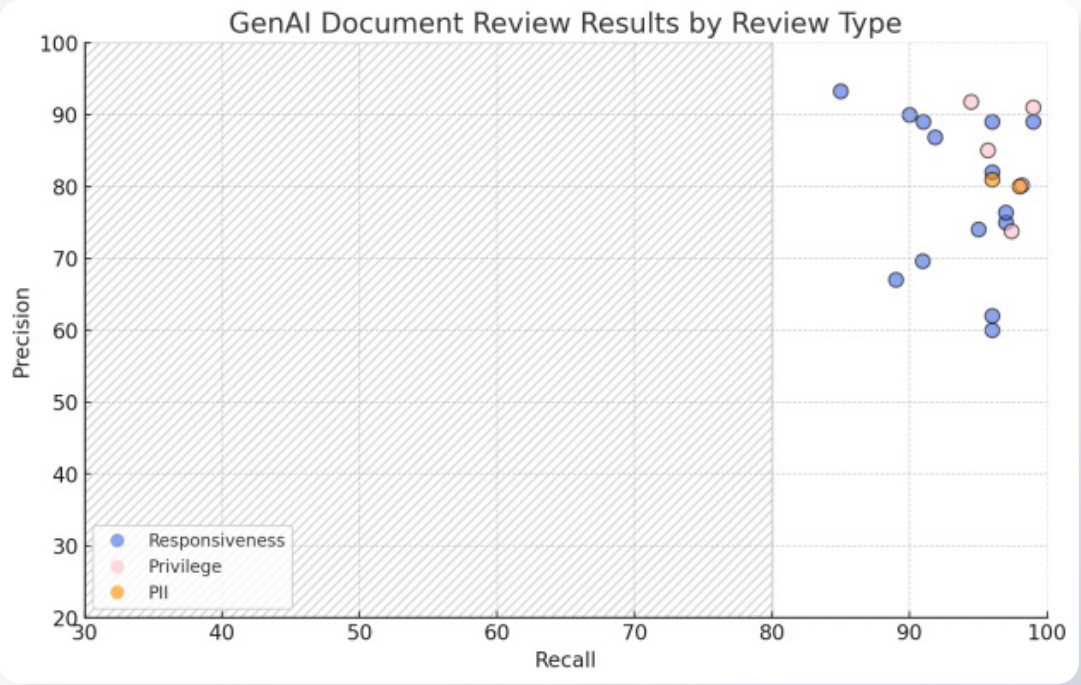
Here, I've added a shaded area representing non-passing results (assuming an 80% minimum for recall).



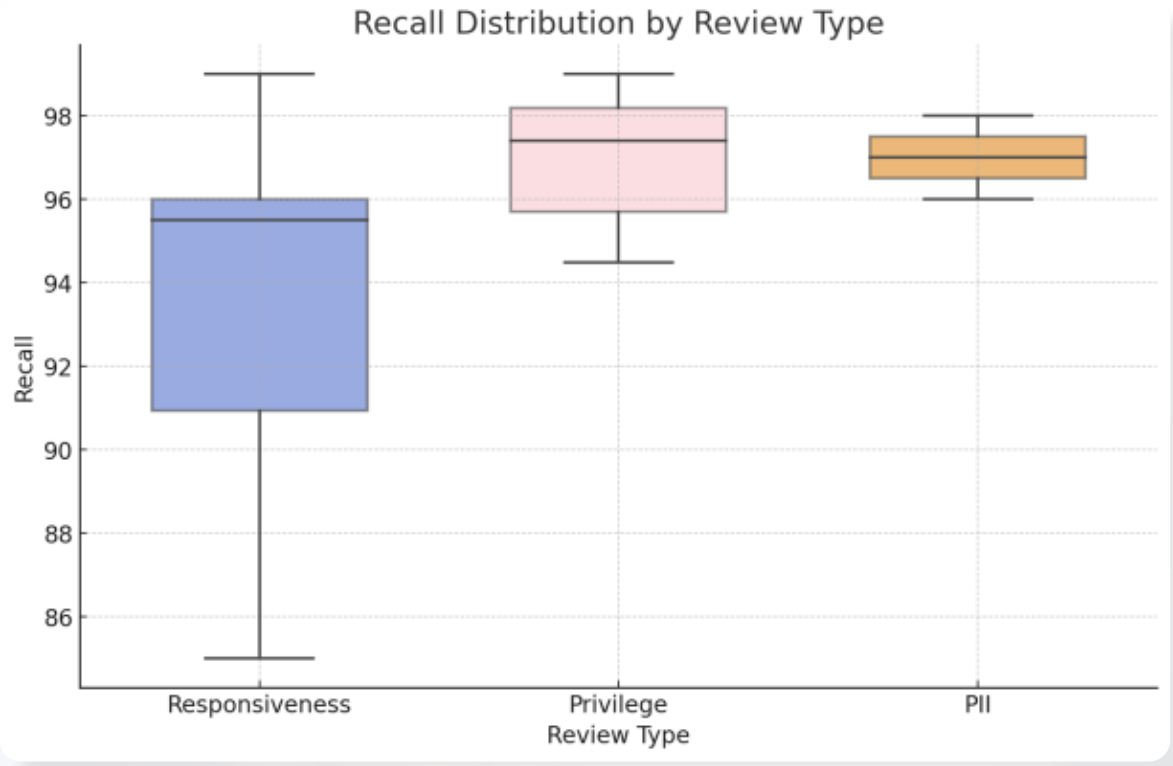
The results show consistently high recall across all projects. Precision values were mostly very good, but showed greater variability than recall, which could reflect differences in user prompting, platform implementation, or the clarity—or ambiguity—of the legal issues in each case.

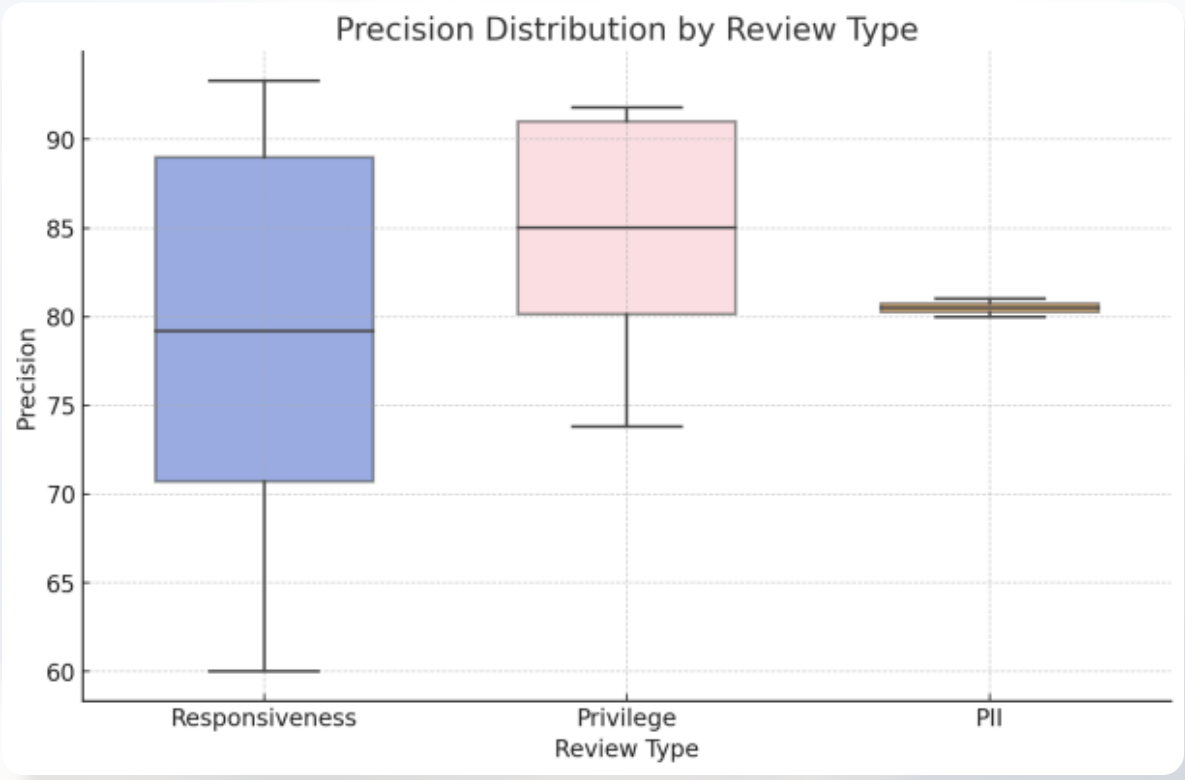
RESULTS BY REVIEW TYPE

One of the pieces of data I was able to gather was the **review type**. We had three different review types: responsiveness, privilege and PII detection (colorized below).



Below are box plots showing **recall and precision by review type**.

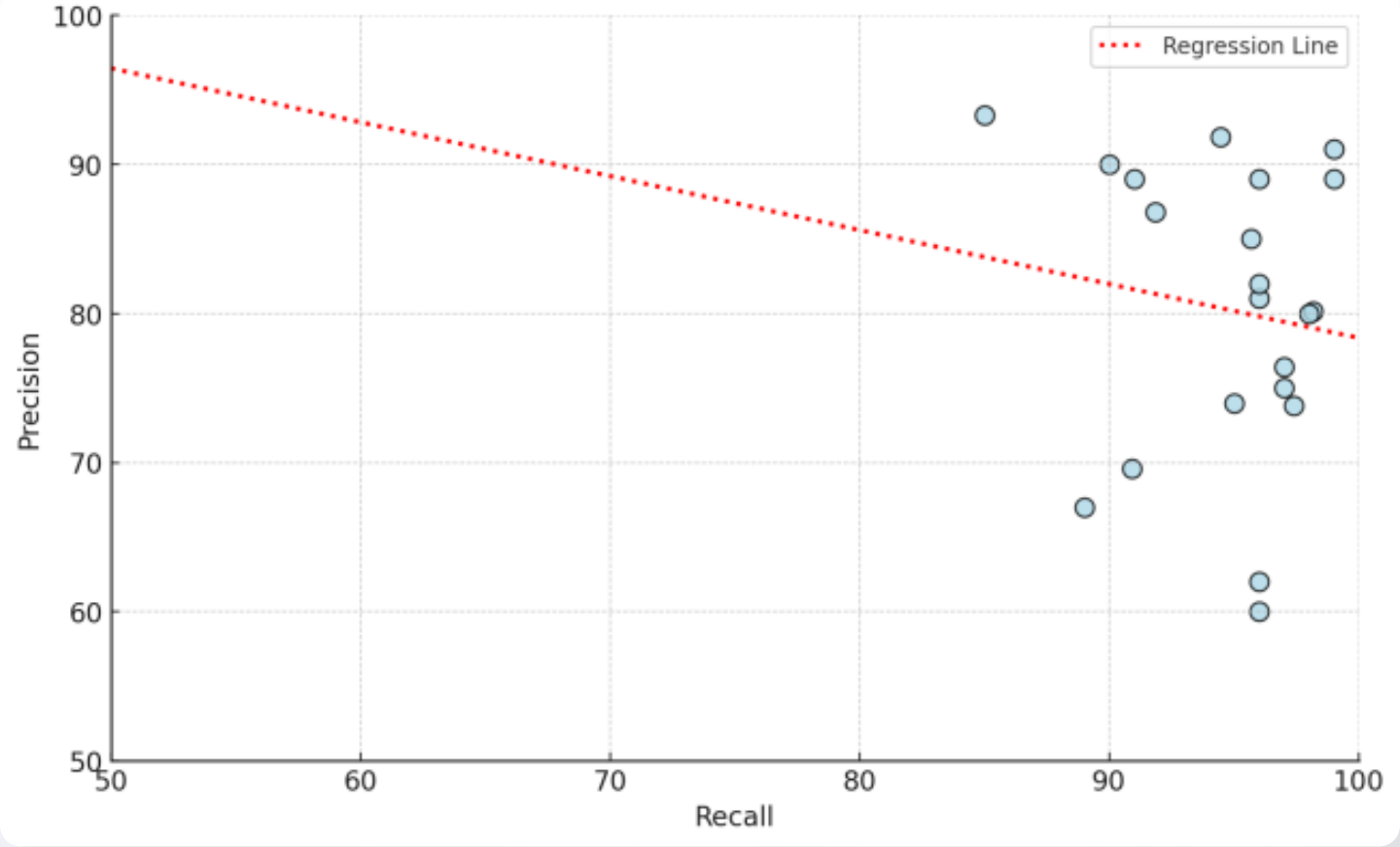




While we observed strong results across all review types, we can see consistently high results for privilege and PII reviews, where review instructions are often more "black and white" than with relevance or responsiveness. Note that two data points for PII is too small draw any conclusions from, but the strong results track with our experience.

VISUALIZING THE RECALL-PRECISION TRADEOFF

In most technology-assisted review platforms—like in human review—there is a tradeoff between **recall** and **precision**. At a certain point, retrieving more relevant documents can only be done at the price of pulling in non-relevant documents. On the flip side, eliminating all non-relevant documents might also mean losing some of the relevant documents. This tradeoff can be visualized with the red line below.



KEY TAKEAWAYS

1. Across a diverse range of project types and practitioners, GenAI-based review performed exceptionally well, easily passing the 80% “recall test” commonly used to validate TAR projects.

2. GenAI performs quite well on all review types tested, including responsiveness, privilege and PII detection.

3. While most projects achieved both high recall and high precision, users have the ability to choose which they optimize for when they design the prompt. See this article for tips on prompting for recall and this one on prompting for precision.

CONCLUSION

My goal with this article is to provide insight into what professionals are observing in the field and to highlight the capabilities of GenAI. Quoting Grace Hopper, *"One accurate measurement is worth a thousand expert opinions."*

Across these twenty-one results, GenAI-based review achieved an average recall of 94% and a median recall of 96%, with all data points easily surpassing the 80% 'recall test' and outperforming both eyes-on review and traditional TAR processes driven by machine learning.

Thank you to everyone who generously contributed data for this study. My hope is that it gives readers a first-person perspective on how the technology performs in the field.

BEN SEXTON,
SVP of Innovation and Strategy, JND eDiscovery