

ADVANCED AI TECHNOLOGY



TAR and Gen AI.

How They Compare

eDISCOVERY AI

Tony Reichenberger



Tony Reichenberger

Director, Analytics Consulting
eDiscovery AI

I've been involved with eDiscovery for over 20 years now and have been involved with TAR and advanced analytics for over 17 years. In that time, I've seen a lot of evolution in the eDiscovery world from paper and early database use, to search terms, to early forms of analytics like concept clustering, to Technology Assisted Review (TAR), and today to Gen AI. I've worked on hundreds of TAR projects or more over the past 15 years, from small

CAL reviews to large TAR 1.0 second requests. My experiences led me and Jim Sullivan to author the book "The Book on Predictive Coding," to help clients and would-be users weave their way through the TAR process. Today, we are going to see another huge leap forward in technology and processes that will impact how we do things, as Gen AI features and workflows get incorporated into day-to-day workflows.

However, one of the questions I'm asked most often is how Gen AI compares to TAR—specifically in terms of quality and workflow. Understanding the differences in how they work, what their inputs and outputs look like, and where they are similar or differ can go a long way toward helping explain why people in eDiscovery are excited about how Gen AI can assist.

The Paleo-eDiscovery Era: Pre-TAR Human Reviews

Set of Documents > Classify Documents > Produce

To many eDiscovery practitioners, talking about the days when we used bankers' boxes, highlighters and redaction tape to conduct legal reviews seems quaint and nostalgic now, and is akin to talking about the Stone Age. In fact, people look at that period in the long, long ago (only 20-25 years) and wonder how we ever got as much done. The answers are less interesting: there were a lot fewer physical copies of material, emails, and other documents back then, and you would just throw A LOT of attorneys at reviewing what was there. The general workflow of a review project could simply be described as 1) reviewers receive instructions on how to classify documents, 2) reviewer gets a batch of documents, and 3) reviewer classifies documents for production.

The only reason I bring it up here is because it's nice to establish a baseline as to where it started, but more importantly, because many of the workflows and habits in legal review that we still abide by got their start from this process. For instance, attachments flowed through the review workflow with their parents and were treated as the same document.

When databases came around, that practice was still common, with family coding (coding every document in a family the same) being the norm. As searches and various analytics features started being used, it became easier, more efficient, and more accurate to code individually on the face of the document and then group families after the fact.

This is an important point because it demonstrates these workflows are not and have never been static; they evolved with each new piece of innovation, just as they do today. Relying on old ways of what works is one of the reasons attorneys are so resistant to change, as they are usually behind the curve regarding technology. Nonetheless, it's also a huge reason why those that adapt early and understand the advantages technology adds can attract more clients and succeed in the marketplace.

TAR: A History and How it has Evolved

Determine TAR Set> Index> Train Model> Classify Documents> Validate> Produce

TAR came into the eDiscovery market in the late 2000s and early 2010s. TAR uses machine learning to analyze the documents you've already coded and suggest other likely relevant materials. To do that, the process requires adequately training a machine learning model to identify what was relevant. An often-used analogy is the way Netflix suggests movies for you based on what you viewed in the past. The advantage of using TAR is that instead of looking for relevant material randomly, you can use the technology to specifically target relevant materials more efficiently, discarding most of the irrelevant material from even needing review.

Various versions of TAR have existed before 2010, but format, structure, and how to integrate the functionality into common legal review workflows initially caused problems. In addition, nobody at the time understood the discovery standards necessary to be considered adequate. Eventually, the Da Silva Moore case in 2012 provided some clarity on those standards. Regardless, there were still some necessary workflow changes that required adjustment.

- ➔ First, not every document works effectively with TAR. Since TAR works on the extracted text of the document, documents lacking extracted text (e.g. image documents, audio files, etc.) or having poor extracted text (e.g. scanned documents) cannot truly be used effectively by TAR. A usual step in the process is to remove from TAR various documents with file extensions indicative of audio, video, image, program or container files, or other non-substantive "junk" files from a TAR review process.
- ➔ Small text volume can be problematic. Individual chats, for instance, may work fine with TAR in theory, but usually there is not enough content there for any model to effectively categorize them. Eventually, a process was adopted to block many of these into specific time periods (usually 24 hours), adding another layer to the workflow
- ➔ Most TAR models require indexing the overall TAR set, which is an important step in the process for the model to make requisite comparisons and contrasts. In addition, any time new documents were added to the set, those had to be included in the index as well, which could take some time to update.

- Documents must be coded individually. It is the only way to adequately train the model and accurately validate the results. The movement to do this began with search and early analytics processes, but there were always some who clung to family coding. TAR put an end to most of that, and now most every review is by individual document “four corners” review.
- Validation is a huge and important function whenever you use technology to cut out documents otherwise needing review. To be defensible, you need to prove that you captured a sufficient percentage of the relevant population, and that further review would be disproportionate in terms of time, cost, and effort to the probative value of any additional relevant documents found. Early TAR features in platforms often had validation as a necessary step in the process, prior to being able to move forward. There are various ways to validate a TAR review, but most all require a random sample in some capacity, whether as a control set (across the entire population of documents) or the elusion set (the unreviewed portion of what remains when you consider stopping review). In any event, the need to do a random sample of documents to get the necessary recall and precision metrics to validate a review is an important step that must be considered and completed.
- Training the model is the most important step in the process. Ensuring that the suggesting model understands all the relevant issues of a matter and is adequately identifying documents for those issues takes a considerable amount of time. Depending on the size of the matter and the number of different issues involved, that could result in training thousands of documents just to get the model performing at an acceptable level. Over time, how one trained the model came to define the type of TAR review that was being used:
 - **Conventional TAR Training (often referred to as “TAR 1.0”):** This was what most people originally thought TAR would be. At the beginning of the project, one would complete the random sample to get metrics and monitor progress and then would continue training until they built up the suggesting model to an adequate level. They would then accept those suggestions (and family members) to produce and then cut out everything else as irrelevant.
 - **Continuous Active Learning (CAL) or Prioritization Reviews:** A CAL review has the benefit of not having to change workflow from a basic human review process. Reviewers simply code documents as they normally would, and the model escalates the most likely documents to the reviewers in their next batch. They continue until no further relevant material is found and then complete an elusion sample to confirm that no (or very few) relevant material remains. Also, all documents that are being produced have received an eyes-on confirmation, helping reassure risk averse attorneys.

At a minimum, CAL became the norm to most reviews simply because it required practically no necessary changes to what you were already doing. One could load documents, batch by family, and send them to the reviewers. From a reviewer’s perspective, there was no noticeable change other than the prevalence within each batch. Relevant documents were pushed to them, and the only additional steps to manage on a day-to-day basis were to update the machine learning (which on some platforms was automated) and monitor the percentage of relevant material regularly as it declined, until nothing else was found. There really was no disadvantage to doing a CAL review this way.

Gen AI: How it Improves Process and Quality

Prompt Testing> Validation> Classify Documents> Produce

If TAR is like Netflix making suggestions based on past decisions, Gen AI is much more akin to Google. You ask for what you want, and it finds it for you. This streamlines the process back towards the original workflow process of providing instruction to the reviewer and classifying the documents. However, to understand the primary differences between TAR and Gen AI, you need to understand a little bit about what the individual underlying models are actually doing.

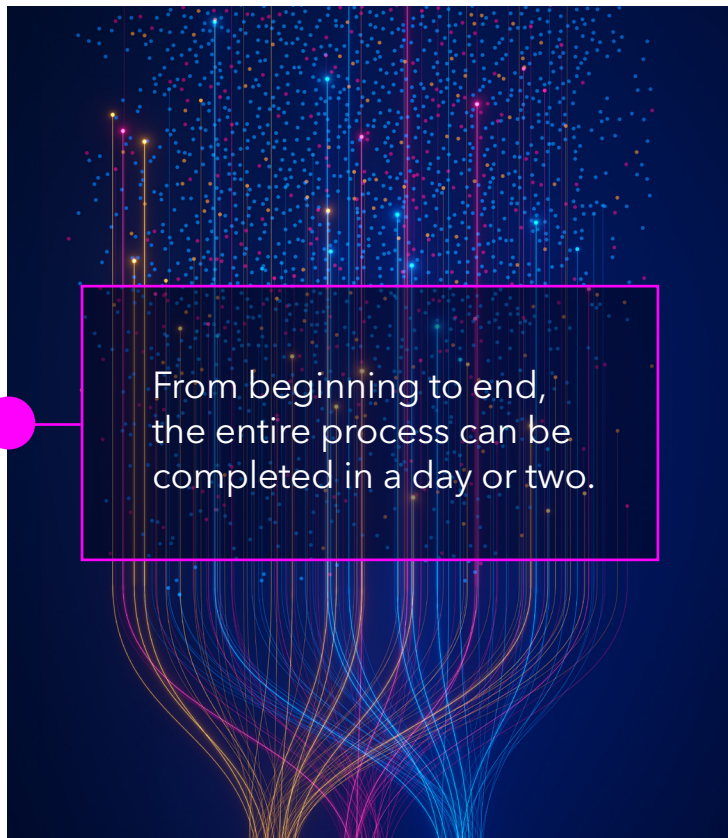
With TAR, you are giving the model one document with all its text and the coding decisions, and another document with all its text and coding decisions repeatedly. The machine learning algorithm (usually a logistical regression algorithm or a support vector machine) draws contrasts between text in the documents you found relevant and those that you found irrelevant. It then analyzes which words and phrases are most indicative of being relevant and suggests those. To be sure, I'm underselling it quite a bit; these algorithms are exceptionally detailed and complex and work in different ways, but in a nutshell, that's what's happening.

With Gen AI, the underlying models are extremely complicated Large Language Models. These models understand language (often multiple languages), context, subtext, grammar, emoticons, synonyms, sarcasm, and abstract aspects of how people speak. This allows for you to make queries of datasets in plain language; you can simply tell it what you are looking for, and it will provide it. This is a much more direct approach than with TAR; TAR models are drawing assumptions based on the text and coding of a document aligning to what you've previously determined were relevant, but Gen AI is basically taking your explicit instructions and using those to determine what is relevant.

This difference can best be demonstrated when remedial training must be completed with TAR. If the model is missing relevant material or is over-suggesting false positives, additional training documents must be found and trained upon for the system to make those adjustments. Quite often the model may not work to make adequate changes, no matter how many additional documents you train, without detrimental impacts to other sets of documents you want to include as relevant. It's a cumbersome, time-consuming, balancing act. With Gen AI, all you need to do is change the prompt instruction and assess the results. LLMs can also provide summaries of documents and insights between documents to help give attorneys an overview of the content within their datasets.

This means that using Gen AI relies heavily on the instructions that you give it, meaning prompt engineering is a primary concern. This isn't much different from what human reviewers have always done—adjusting their approach based on evolving instructions as they progress through document sets. How one describes what is looked for, organizes the instruction, and provides it to the LLM, can have a large impact on the results one receives. And from experience, most mistakes in output can be traced back to how the prompts are worded and what is being asked; the LLMs are quite literal in this regard.

For this reason, you'll need to test your prompts to ensure they effectively identify the documents you want while filtering out those that are irrelevant. In many respects, this part of the process is similar to the training phase of a TAR 1.0 project; however, it is much easier and quicker to complete.



From beginning to end,
the entire process can be
completed in a day or two.

After testing the prompts on a small set, once you are satisfied with the results, they should be validated with a control set random sample. This means taking a 95% confidence / 3% Margin of Error sample (or other size sample), the same way you would with TAR, to get recall and precision metrics measuring the effectiveness of the prompts and AI. On most TAR projects, the recall goal is usually 70%-75%, and additional training and model adjustment are conducted throughout the process to either achieve or maintain that recall threshold. Unfortunately, there is always a trade-off with precision; if the TAR algorithm suggests more broadly, recall may go up, but at the expense of precision and adding more false positives into the production set. Attaining a high recall while maintaining relatively high precision, particularly while additional training is ongoing, is a non-stop consideration in TAR use. With AI, most projects regularly achieve 85-90% recall and 95% recall is not unheard of. Precision

with AI is also notably better since the prompts are direct and can be directly refined by additional instructions to carve out irrelevant false positives. Defensible recall levels can be accomplished much quicker, easier, and simpler with Gen AI than with TAR, while simultaneously achieving high precision.

Once you have validated the prompts and achieved the level of recall you want, you can apply the prompts broadly across the dataset. Because AI is much faster than human review and applies prompts consistently across every issue and document—regardless of its position in the dataset—the results are significantly more accurate than those from human review or CAL.

From beginning to end, the entire process can be completed in a day or two, with the size of the dataset being the greatest variable due to processing time, just as it is with TAR. But where TAR requires indexing the set to run the machine learning to compare documents, Gen AI requires no such indexing, making the process more scalable and faster beginning to end. This also means that when additional documents are added to a TAR database, they must be included in the index and trained into the model, which may require additional training to re-attain the recall threshold. With a Gen AI review, it's just a matter of using the existing prompts against the new data and pushing the button to submit.

| Assuming 1,000,000 document review | Human Review (50 docs/hr) | Technology Assisted Review | AI Review |
|------------------------------------|--|--|--|
| Hours to Complete | 20,000 hrs | A few hours to train, machine iteration may take an hour. Multiple iterations will be necessary to get suggestions adequate. | A work day |
| Add 10% Human QC | Additional 2,000 hrs to look at the set eyes on. | See Human Review; TAR is able to identify discrepancies b/w human and TAR to target | See human review; AI review provides explanation of coding decision to consider. |
| Time to set up | Little. Time to batch documents and delegate to reviewers | Substantial; must sort and exclude docs that cannot be sent to TAR; must index docs; must create training sets and train documents | Little. Just time to enter prompts. |
| Indexing required? | No | Yes | No |
| Validations | Typically, Human QC of various samples, either by random set or judgmental sample set to various parameters. | Control Set Elusion Sample Various Human QCs | Control Set/Elusion Sample Various samples of AI output Explanation of coding decision |
| Average Recall/Precision | ~65-75%. Precision is equal to the prevalence of the set (lowest it can be). | Flexible, usually targeted to 70-80% recall, with varying levels of precision depending on set and training. | Regularly recall over 90%, with precision above 70 or 80%. |

| Assuming 1,000,000 document review | Human Review (50 docs/hr) | Technology Assisted Review | AI Review |
|---|---|---|--|
| Workflows | Linear review of all documents. | CAL review, resulting in linear review by priority until cutoff. TAR 1.0 review, resulting in sample training to improve model and cutoff at predetermined recall score. | Can code all the documents in single setting or be used in conjunction with other database features. |
| Advantages | Every document is reviewed eyes on; offers assurance of quality (even though reviewers are correct only 75% of the time). | Can remove irrelevant documents from review resulting in substantial cost savings. | Fast, efficient, more accurate and much easier to manage. |
| Relevance Adjustments During process | Must re-review all impacted documents | Must retrain classifier to adjust to issue | Can apply prompt revision or addition quickly. |
| Documents Added During Process | More human review | Must be included in the index, additional training to accommodate new documents/ issues; must re-validate for metrics | Can apply existing prompts to new set; will have to re-validate for metrics. |

Comparing Human Review, TAR and Gen AI Directly

Many want to genuinely see how each review workflow works and compare them to one another, and there is a very easy way to demonstrate how.

First, take a case in which a human review was conducted. Preferably, this would be something with 20,000+ or more, with multiple relevant issues associated with it. Because you already completed the review, you know how long it took you to review it, the costs involved, the average review rate and other statistics that can be used for comparison. This is your baseline. You can also take a 95% confidence / 3% MOE sample (~1,065 documents) from this set as a control set to use when validating TAR and Gen AI.

Next, rerun the project as a CAL project using the existing coding. You'll need to create a new field for the machine learning model to use for training. As documents appear in the process, simply mass code the new field for each batch based on the original coding. Submit 500 documents at a time, retrain, code the next 500 documents and reiterate until you get sets with very low prevalence. Once this is complete, you can assess the recall, precision, and accuracy accordingly, and estimate the amount of time based on the number of documents reviewed as part of the CAL process and the review rate assumptions from the human review.

Lastly, to compare Gen AI, you can take the original instructions for the human review and put them in as prompts. Run them against 500 documents in the set (not within the random sample) just to make sure they return what you want and don't need to be revised to capture more relevant material, or alternatively, carve out false positives. Once comfortable, you can minimize costs for this test by running Gen AI only on the random sample (for instance, eDiscovery AI will cover a small set of documents for you to test it on as part of a Proof of Concept); the amount of time to enter in prompts, whether 1000 documents or 1,000,000, remains the same. This will give you the effectiveness metrics for comparison to the other sets. Cost differences are usually based on a per document basis, so they can be extrapolated across the size of the total set if need be.

This process will give you an apples-to-apples-to-apples comparison of quality, time, and cost among the three different workflows.

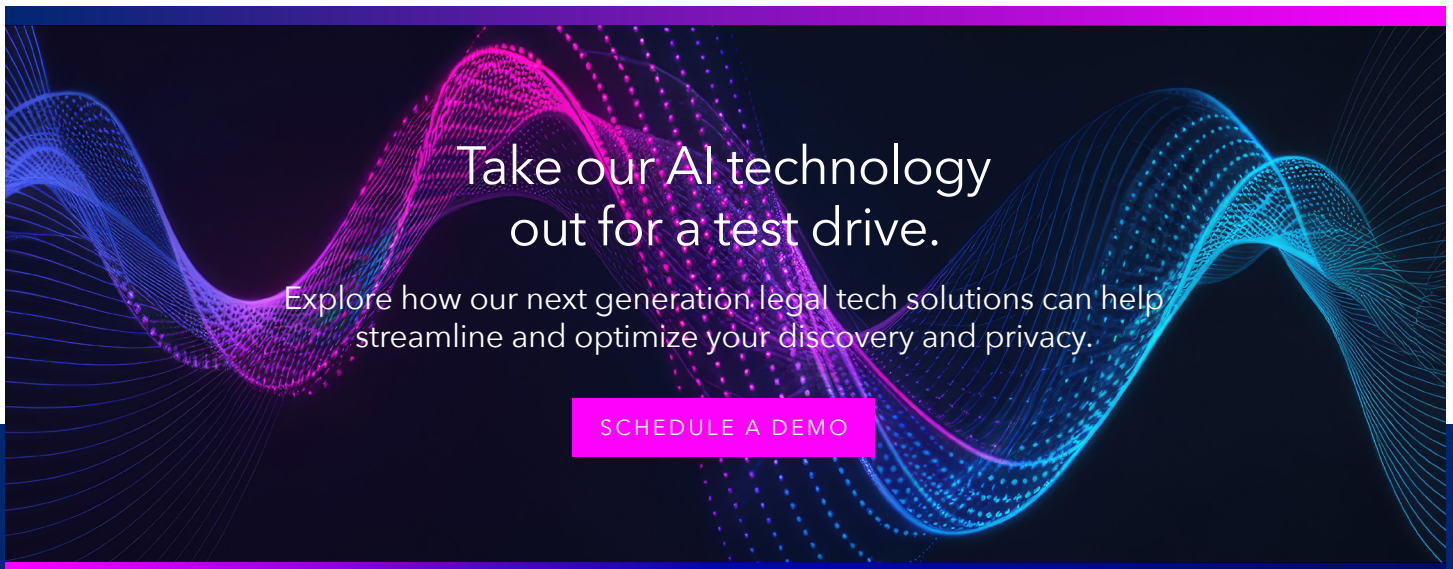
Conclusion

Legal review processes are not static; they are ever evolving to take advantage of technological innovation and remove inefficiencies from the typical workflow. The giant leap forward with Gen AI is no different in this regard. As Gen AI continues to increase in adoption, we can anticipate further changes to the workflow as more users discover huge advantages and cost savings in how they manage AI reviews. Nonetheless, Gen AI is here to stay and the sooner one adapts to the new workflow paradigms and seeks out ways to improve them, the better their business and market advantages will be in the future.



Welcome to **eDiscovery AI**

eDiscovery AI is a data intelligence company building the next generation of legal tech solutions. We provide AI-powered solutions to streamline and optimize legal discovery and privacy. Our suites of solutions - Early Case Intelligence™, Review, Privacy, and Multimedia - provide industry leading features, speed and accuracy. We are dedicated to delivering advanced AI technology with expert guidance to help our partners navigate the evolving landscape of legal technology.



Take our AI technology
out for a test drive.

Explore how our next generation legal tech solutions can help
streamline and optimize your discovery and privacy.

[SCHEDULE A DEMO](#)

eDISCOVERY AI