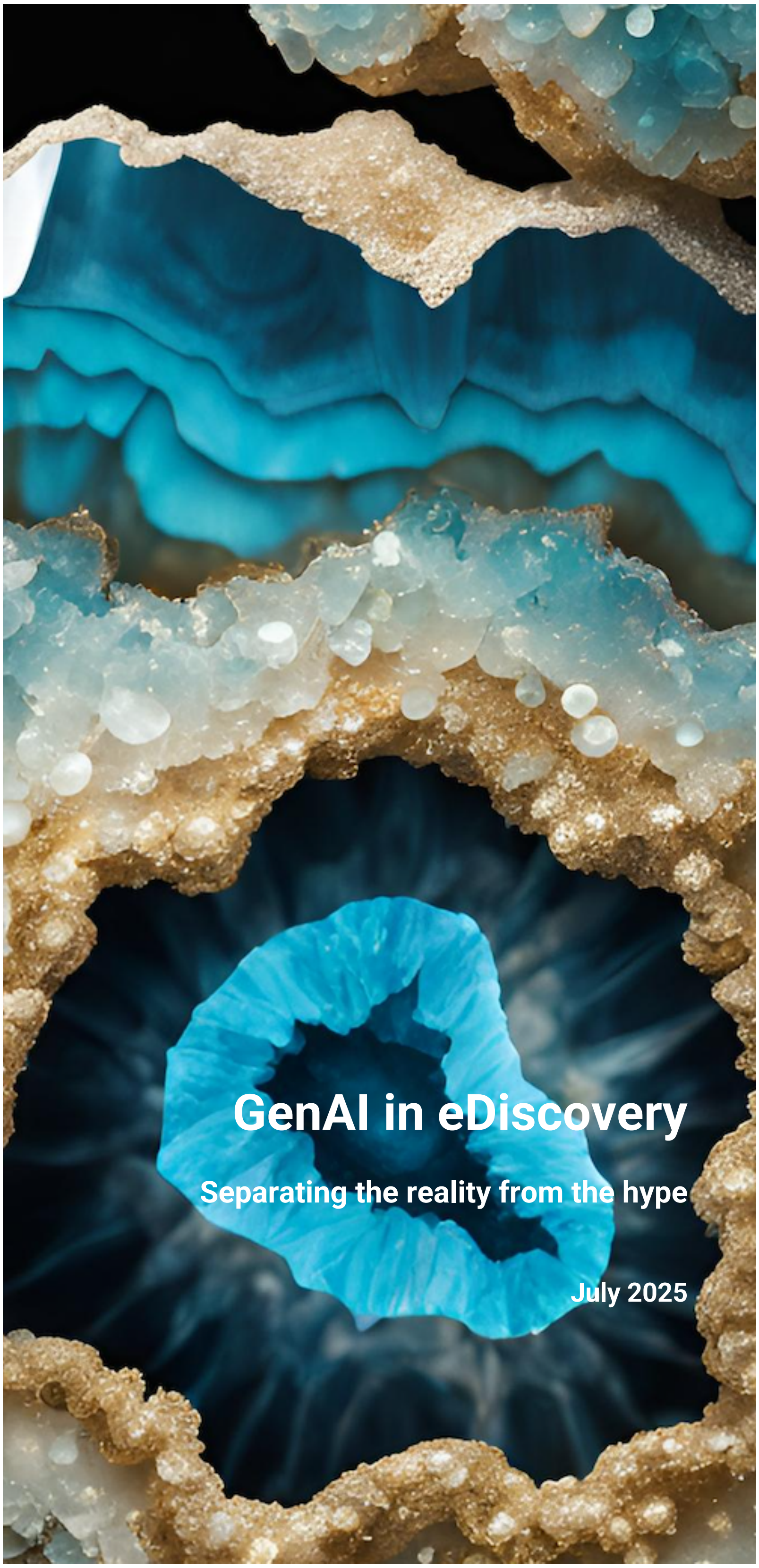
salient e-discovery

GenAI in eDiscovery gets real

Practical use of GenAI



Introduction

Generative AI (or GenAI) is seemingly all around us and depending on your philosophical position, is either a good or a worrying thing.

Conventional AI has been in use in the eDiscovery world for quite some time now in the form of Machine Learned models such as in TAR (Technology Assisted Review) and CAL (Continuous Active Learning, or TAR 2.0). Or indeed the unsupervised learning models such as entity extraction and other Natural Language Processing techniques like conceptual searches and clustering that proliferate in all sorts of products and services in everyday life, quite aside from eDiscovery.

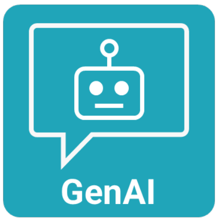
GenAI, on the other hand, feels that bit closer to true sentience than Conventional AI, apparently passing the Turing Test with ease through the synthesis of responses and positions to the questions posed. And as such, perhaps it causes more concerns that the technology will replace the human. But with much press coverage of plagiarism, bias and of course hallucination (a euphemistic way of saying 'making things up'!), there's probably a way to go before we should be truly concerned.

In the meantime, the technology ***can*** be harnessed to good effect but does require users to see it as a helpful tool rather than purely a threat and as such, we are probably obligated to learn how to best use it.

Computers have always been far superior to humans at churning through large volumes of material without tiring or needing a break, and of assimilating data across far larger volumes than the human brain can reasonably expect to do. With the ever-increasing volumes of electronic evidence in a typical eDiscovery matter, it would seem logical to leverage the power of computers to assist, so how well does GenAI meet that challenge?

This article looks at the evolution of Gen AI in eDiscovery, from the basic concepts to an exploration of the strengths, weaknesses and considerations when considering it for your eDiscovery workflows.

Gen AI – the Basics



But let's start by de-mystifying what we mean by GenAI with a high-level overview of the process that is most prevalent in both eDiscovery applications and the more familiar Chat GPT and CoPilot type solutions.

At the risk of using a few technical terms, the generic approach used is called ***Retrieval Augmented Generation (RAG)*** or put more simply, a search-based approach to the synthesis of responses. After all, what is eDiscovery if not a search exercise?

RAG uses a range (and often a combination) of search techniques to find content that is most likely to be relevant to the question that has been posed, and by doing so provides ***context*** to the so-called ***Large Language Model (LLM)***, of which there are many on the market, including Chat GPT, Claude and GROK.

The question that you ask, the ***prompt***, in conjunction with the ***context*** provided, ***grounds*** the LLM to focus the LLM's generic 'understanding' of the topic and apply it / make it relevant to the context you are providing it.

Large Language Models are trained on extensive amounts of content and, as such, have what presents as an apparent 'understanding' of many subjects. In reality they have absolutely no sentience at all – they are essentially used for predicting what the next word should be when generating a response – but based on that extensive training set, coupled with the advanced algorithms used, the effect is an apparent understanding.

A major concern that is often discussed in the media concerns data protection, particularly in the context of the legal industry. By asking questions of your content, are you inadvertently releasing your material (or worse, your clients' material) to the public domain and training the models further? In this respect, choose your GenAI solution wisely. Most, if not all GenAI solutions serving the legal sector are hugely aware of this issue and as such are ***closed systems***. With these solutions, no data passes into the LLM; think of it as only looking at your content through its impenetrable glass ***context*** window and applying its understanding to that, ***grounded*** by the question asked.

Finally, the GenAI solution will synthesise its response and reference the evidence that it has used from your case material, to construct its answer.

This is an over-simplification of the actual processes, but it's sufficient for explaining the generic approach.

Prompt engineering

Another topic that you may have heard discussed is Prompt Engineering. Or in laypersons' terms, the 'art' of how you construct the question you send to the GenAI solution.

This can be a daunting topic and there are now many courses available covering the subject.

Meanwhile, vendors are putting considerable effort into their solutions to remove or reduce the impact of prompt quality, to allow users to ask questions as naturally as they would speak to their colleagues. The science behind it is again Natural Language Processing (NLP) but to all intents and purposes is intended to help interpret the intent or the contextual meaning of the provided question, so that it doesn't have to rely purely on the lexical terms in the prompt for conducting its search.

However, it still pays to consider how you construct your query. Getting the prompt right is critical. And to do that, we need to scratch a little further beneath the covers to understand the components of the prompt and their relationship.

You can break your input down into two key components; the ***question*** and the ***instructions***.

The ***question*** is just that; the question to which you are seeking an answer. Behind the scenes, this may well be further manipulated by the particular solution you are using to infer meaning or refine the searches run, based on the context you provide.

The ***instructions*** provide the solution with guidance as to how you want the response structured when it synthesises its response. For example, do you want a summary, or detail, a report format, a tabular response or indeed an answer in a different language or even written as a poem!

The actual prompt is typically a long, proprietary and internal device for each solution, which uses various inputs, not least of which are the question and instructions that you provide.

But having a better understanding of the effect of these 2 elements is critical in constructing meaningful queries and hence in receiving optimal results from the GenAI component in your eDiscovery solution.

So, what does each of the inputs do in more detail? You need to think in terms of Search first and then LLM Reasoning second.

The Question - Search

Competent eDiscovery systems which have GenAI components will typically build a more comprehensive index of the case content, often called a semantic index. These use AI models themselves to understand the meaning and context of data, rather than rely on simple keyword matching.

It is the key terms and phrases used in your question which will be used to filter results from this semantic index. It is therefore critical to get that right, to ensure the most relevant elements of your content are selected.

When you submit a question, the query that is run against the semantic index will use the terms and phrases provided quite literally in the question, to identify the likely most relevant content to pass to the synthesis of the response.

Instructions – LLM Reasoning

Instructions provide the LLM with guidance as to how it should construct its synthesised response. It has no influence on the content which is selected from the semantic index from which the response will be built, as it is applied *after* content has been returned from the search.

Accordingly, getting the question 'right' can have significant benefits. As a general rule of thumb, focus on the ***query design*** for the question aspect and then ***prioritisation and structure*** of the response in the instructions.

However, there is no absolute right or wrong approach and a degree of experimentation will be necessary, but consider the following factors when building your queries:

1. Explicitly include the key topics or themes of the query in your question rather than being generic.
2. Consider the complexity of the query. Rather than try to build too many aspects into the desired response, break the query up into smaller tasks. This will reduce the likelihood of hallucination or confusion in the response as the semantic search task that gets run will be more focused.
3. Judicious use of terms like 'top', 'key' and 'most'. The use of terms such as these will likely be interpreted in the search phase of the process ***quite literally***, and as such, are not necessarily valuable. Not all GenAI engines are the same. Some will be tuned to try to provide exhaustive responses, and others will have set limits to improve response performance. If the latter, the engine will not provide an exhaustive list. However, it will ***always*** try to return the 'top' results, irrespective of the instruction, as it returns the most semantically relevant content before sending the results for synthesis of an answer.
4. Granularity. Remember that solutions may have an enforced limit for results. If the context has a lot of relevant content, your response may exclude less highly ranked but still relevant content. Keep an eye on the number of items used in generating the response (most solutions will report this) and if it has hit the limit, consider making your question more precise and granular.
5. Instructions. These can be used to influence the format of the synthesized response. For example, you can request the response in a report format with a summary, presented in a chronological sequence, or perhaps as interrogative questions that may be used in interviews.
6. Chronology. If you are looking for timelines in your responses, it should be noted that these are usually derived from the text of the content, not from any associated metadata. If you want to restrict searches and questions to a specific date range based on metadata, you should use conventional date filtering features to pre-select content and then ask your question of that corpus.

The Scope of GenAI in eDiscovery

Whilst there are several areas where Gen AI could be effectively applied in the eDiscovery workflow, we explore the 3 primary areas of ***Early Case Assessment*** (ECA), ***Review*** and ***Document-level GenAI*** in the following sections.

GenAI for ECA

With the ever-increasing volumes of electronic data having a costly impact on eDiscovery, the first area where GenAI can be applied is in Early Case Assessment. Specifically, being able to assist with the early assessment of large evidence estates, to confirm the fact patterns as we know them and to uncover further insights; the other 'unknown unknowns', so to speak. The value proposition being to accelerate the overall investigatory journey by focusing in on the relevant evidence at an early stage and thereby cutting the overall time and hence cost for the exercise.

The single most expensive aspect of any eDiscovery exercise remains human review. "Eyes on" legal time. So, the objective has long been to cull volume as early as possible in the process, however by doing so, the risk is that you cull valuable or pertinent evidence. Finding the optimal position is therefore crucial.

It's also the case that we often don't know what we don't know when embarking on these exercises. And given that disclosure tends to still be based around key words or phrases, making sure you are best placed to agree beneficial terms for disclosure is also key.

So, the strategy with using Conventional AI and Gen AI, allows you to cast the net that bit wider initially, then to enrich and mine the resultant collection intelligently, to allow you to reduce the volumes that go anywhere near the document review stage, with far more confidence.

Key to a successful strategy is to use the tooling to support the investigative mind set – not expect it to be a magic wand. Explore the themes surrounding a case, just as you would ordinarily, but now with more power at your fingertips. The GenAI for ECA solution works best when fully integrated into the rest of the platform, so you can use Conventional AI techniques (either in isolation or by combining several techniques) to help you make your queries and the context you ultimately send to GenAI more granular.

Those conventional techniques might include:

- ***Communications analysis***. For example, you may have filtered around key terms but then use the communications view to further limit the context to the relevant communicators or perhaps just the domains of interest. Or perhaps by analysing communications you may expose further persons of interest that you were not aware of, that warrant deeper scrutiny or inclusion in the investigation.
- ***Conceptual search***. Rather than using pure keywords, you may want to explore other related concepts or terms that occur regularly in proximity to the initial keywords or terms. If your eDiscovery solution supports this, you may also be able to do so graphically. By limiting the exercise purely to keywords, you may miss critical, relevant content that uses synonyms and certainly restrict yourself from exploring those parallel concepts that may also be worthy of consideration.

- ***Clustering***. Cluster wheel technologies, when combined with GenAI, provide an excellent way to cast the net a little wider before culling back to the more relevant content, to improve the likelihood that you have considered the most relevant content in any investigative process. By considering lexically similar terms and commonly related vocabulary to the initial search terms (keywords), you may wish to submit the entire cluster to the GenAI ECA engine, to check if the resultant references include relevant content from the estate that would otherwise have been missed.

GenAI for Review

Whilst the Early Case Assessment application of GenAI is excellent for accelerating the early stages of an investigation, the next logical leverage of GenAI is in supporting the review process.

Despite judicious application of technology to intelligently select likely relevant content, it is inevitable that as data volumes grow, the absolute numbers requiring review will remain high. However, clients still want a rapid response, which will typically necessitate the use of more junior or less experienced resources to wade through the review.

Conventional AI can assist here with Machine Learned (ML) models which learn from the coding decisions made to promote similar content for prioritised review, but that implies that you have experienced reviewers on the case from the very start and furthermore, the time it takes for the ML model to stabilise will vary, depending on many factors including the consistency of responses provided and the richness of the data set in the first place. [By richness, we mean the density of responsive documents in the dataset. If you only encounter one responsive document for every 1,000 reviewed, it will take longer for the model to converge to a point of stability than if it's 20 responsive documents for every 50 reviewed].

What if Gen AI could further assist in accelerating this process?

The Concept, Practical Challenges and Reality

The idea is, quite simply, to use GenAI to make decisions on a document, based on some contextual understanding that the user provides.

However, who is going to trust a machine to have made the right decision? Certainly not the professional indemnity insurers. Whilst it might be technically feasible to conduct a review entirely by machine, the reality is that the machine will only be as good as the instruction it is given, so needs to be well taught and most importantly, backed up by human-in-the-loop quality control.

There are several solutions in the market which use GenAI in various ways to guide the reviewer to a coding decision, in-so-doing providing justifications for the guidance, along with counter positions and other considerations. However, at this point we will focus solely on the solution from Reveal, which is scheduled for release in the second half of 2025.

Definition and Review Support

Reveal's solution takes a very considered and pragmatic path. The first step of the process is to provide the engine with a contextual understanding of the case in hand; what it's about, the known fact patterns that make up the investigation or litigation and the nature of the responses that are requested by court.

The Gen AI Advisor component then analyses the supplied natural language to identify items, entities (people, places, companies, monetary amounts, etc) and concepts which essentially paraphrase and represent its understanding. The investigator or lawyer can interact with the Advisor to adjust the weighting of any particular factor and to finesse the definition of the case.

Think of this as the ***calibration*** process. The definitions (i.e. the prompts which are asked) are the single largest contributor to the end results, so it is critical to have ways of testing these prompts on a smaller, manageable subset of documents. There are multiple ways that this definition / prompt calibration process is integrated into the Reveal review workflow but all allow the user to manually review this smaller calibration universe of documents, the purpose of which being to evaluate the degree of alignment between the LLM decision and the human decision. This can be used by the human ***subject matter expert*** to adjust the definitions, with or without automatic suggestion by the definition advisor process.

Ultimately, a tested, verified definition is a much safer and smarter way to then unleash the LLM review on a larger set of documents.

Having provided this clear definition, GenAI for review in Reveal can then promote documents which are most responsive to these criteria and provide a rationale along with alternative considerations which can help the less experienced reviewer with their coding decisions.

In theory, the system could complete an unsupervised review of the document set itself, based on the provided definitions, but in reality, there are few law firms which would entrust this process entirely to a machine. The more likely use would be to support and accelerate, rather than replace, the human review. Or to consider the tool as a first level robot reviewer.

Hybrid Mode - Accelerated Model Building

The second opportunity for reducing the cost and time in review is to accelerate the time it takes to build conventional Supervised Machine Learned models.

As mentioned previously, the time for a model to reach stability is highly variable. But what if you could use the machine to rapidly review, make decisions and stabilise a model? We may think that '***supervised***' implies by a human, whereas it actually refers to the model learning from a labelled dataset, whether that's human or indeed machine supplied. In ***Hybrid mode***, Gen AI for review pulls out AI batches, decides whether each item is positively or negatively responsive to the prompt, and this continues to iteratively train the ML model.

We're not suggesting that anyone uses the outcome to provide advice without any human review, but it can short-cut the delay before AI generated review queues can be made available to the human reviewers, at which point the Continuous Active Learning (CAL) model would use human responses to validate the degree of correlation between machine and human decision making[1].

Crucially, this ensures that essential human review of the content takes place and if there is a mismatch, CAL kicks in and any inconsistencies in the model are corrected.

Maybe one way to think of it would be as a first level review, before a more experienced consideration of the initially identified evidence.

Document-level GenAI

A third obvious application of GenAI is to provide rapid summaries and helpful analysis at a document level.

Such summarisation is very helpful in gaining an overview of the individual document and outlining the key-points, especially if the document is lengthy. By asking any question at the document level, you limit the likelihood of hallucination to that item by eliminating any cross-document contamination. Of course, the results still require human validation but the technique is potentially very valuable for accelerating the review of a single document and where a contextual response may be more helpful than a keyword search alone.

Conclusion

Whether you're a sceptic or wholeheartedly embrace GenAI, what is certain is that it isn't going away. As ever, the trick is in separating the hype from reality, leveraging where useful and ignoring the irrelevant or fanciful.

If you are ready to embrace the cost and time saving potential of GenAI, get in touch with Salient, where we can demonstrate the value of the GenAI solutions embedded within the Reveal platform that powers our eDiscovery service.

[1] In data science terms, the approach demonstrates transparency and defensibility by combining recall, precision and F1 scores into the Gen AI workflow