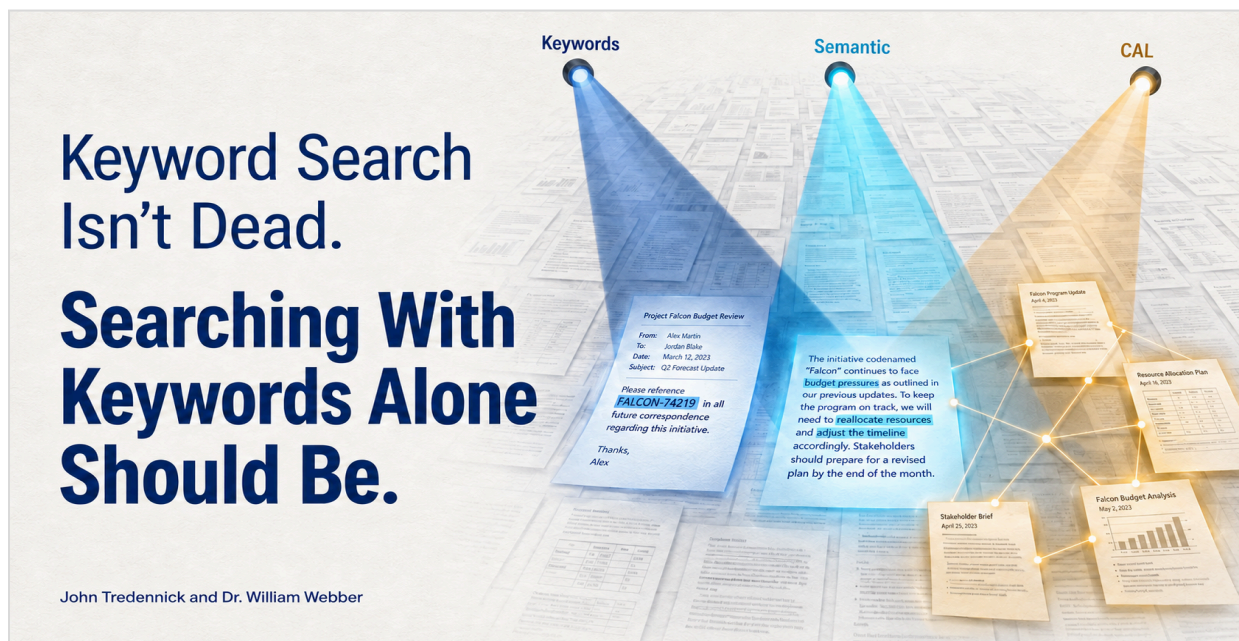




Keyword, semantic search and continuous active learning solve different problems. Used together, they can find evidence that any one method would miss.

Published by EDRM and J D Supra 09-2026

Editor's Note: This is the first article in a four-part series exploring AI-Integrated Search and its role in AI-Integrated Discovery. The full series [can be found here](#).

By John Tredennick and Dr. William Webber

The most dangerous document in a case may be the one nobody knew how to search for. It may describe a contract breach as "the supplier problem." It may refer to a secret project by an undisclosed code name. It may contain a misspelled company name, an unexplained acronym or a seemingly harmless phrase whose significance becomes apparent only months later.

Keyword search will not find that document unless someone anticipates the words it contains.

That does not make keyword search obsolete. It makes keyword search incomplete. Legal teams can now combine it at scale with two methods that look for different signals: semantic search, which matches meaning, and continuous active learning, which learns relevance from reviewer decisions.

“The point is not to replace keyword search. It is to stop asking for one method to solve every retrieval problem.”

The question is no longer which method should replace the others. It is what becomes possible when all three work together.

Three methods, three blind spots

For decades, searching a large document collection meant supplying words and retrieving documents that contained them. The profession made that process increasingly sophisticated with stemming, proximity operators, fuzzy matching and Boolean logic. Modern lexical systems such as BM25 can also rank results, giving greater weight to distinctive terms and accounting for document length.

Keyword search earned its position. When an exact term is the signal, it is usually the best tool available. That includes:

- Statutory citations and quoted contract language.
- Case numbers, account numbers, docket identifiers and document IDs.
- Names, product codes and other distinctive terms.
- Court-ordered or negotiated search terms.

Keyword search is fast, transparent and easy to explain. A judge can understand a term list. That matters.

But keyword search has a structural limit: it finds the words the searcher thought of, spelled the way the searcher expected to find them.

Semantic search works differently. It compares the meaning of the query with the meaning expressed in the documents. A search for failure to perform might surface discussions of "material default," "missed deliverables," "unfulfilled obligations" or "the supplier problem," even when those documents contain none of the words in the query.

Continuous active learning, or CAL, looks for a third kind of signal. As reviewers mark documents relevant or not relevant, a classifier learns from those decisions and ranks the remaining collection. It is not limited to the original query or to a general model of linguistic similarity. It learns what relevance looks like in this collection, for this matter, based on the team's judgments.

Each method also has a characteristic weakness.

Keyword search can miss material expressed in unexpected language. Semantic search can overgeneralize and is generally weaker on exact identifiers. CAL depends on the quality and range of the examples reviewers provide; poor or unrepresentative judgments can send it in the wrong direction.

The failure modes differ enough to make the methods complementary. Methods that fail in the same way are redundant. Methods that fail differently can catch one another's mistakes.

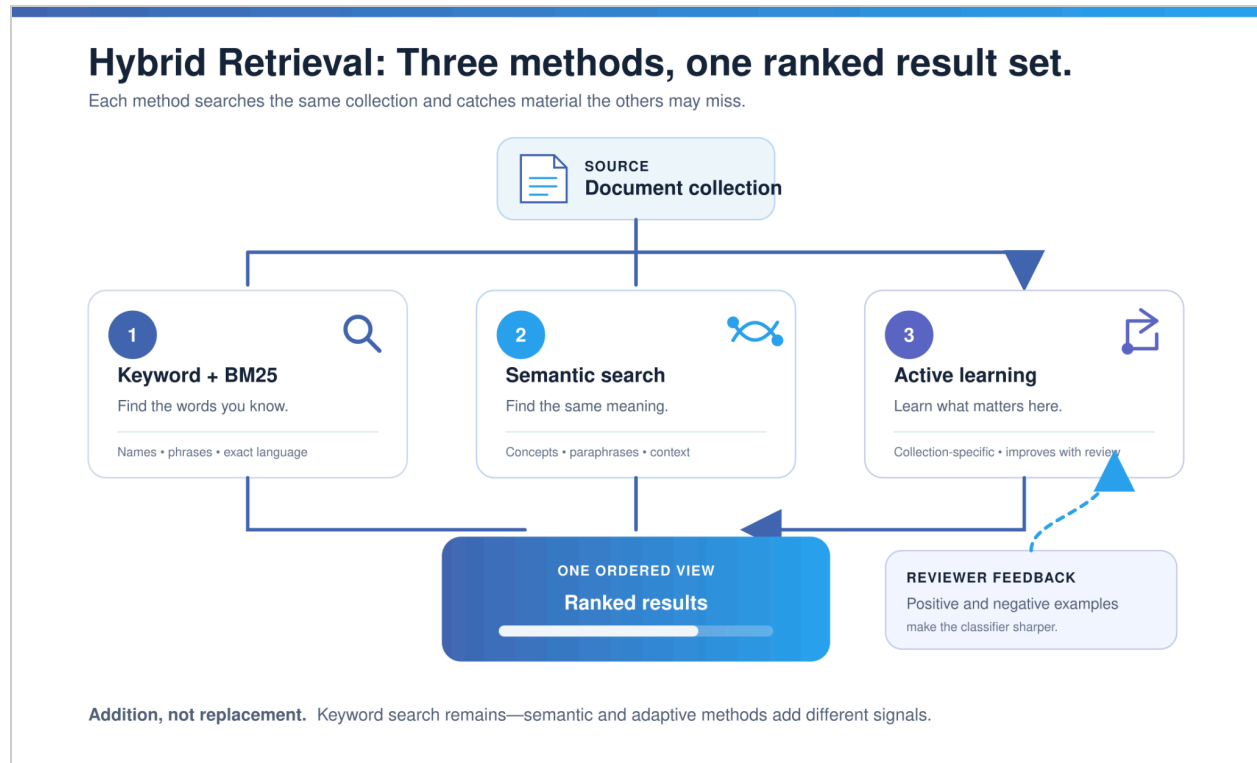


FIGURE 1. THREE METHODS, ONE RANKED RESULT SET Keyword, semantic and continuous active learning contribute different signals; reviewer feedback sharpens the CAL ranking as review proceeds.

The confidence problem

The difficulty with keyword search is not simply that it misses documents. Every search method misses documents. The deeper problem is that keyword search provides no reliable way to see what its vocabulary excluded.

Blair and Maron's landmark 1985 study remains the clearest illustration. Experienced searchers worked with a real litigation collection until they believed they had found roughly 75 percent of the relevant documents. When the results were measured against the full collection, they had found about 20 percent.

The study is old, and its collection was small by current standards. Its enduring lesson is not that every keyword search achieves 20 percent recall. It is that searchers cannot infer recall from the apparent quality of the documents they found. The missing documents are often invisible precisely because they use language the searchers did not anticipate.

Consider a contract dispute. Relevant documents might refer to "breach of contract," "failure to perform," "material default," "nonperformance of obligations" or simply "they didn't do what they promised." To a keyword system, those are different strings. Broaden the query to cover enough

variations and the result set can fill with false positives. Narrow it for precision and relevant material may be systematically excluded.

Nor does reviewing everything create a perfect benchmark. Roitblat, Kershaw and Oot compared two professional review teams with the original review of the same collection. The teams achieved roughly similar recall, but they did not identify the same documents. They agreed on only about 28 percent of the documents either team marked responsive.

Human review remains indispensable, but human relevance judgments vary. There is no pristine set of relevant documents sitting in a collection waiting to be revealed without disagreement. That is why a defensible process should not be judged against an imaginary standard of perfection. It should be judged by how well it searches, how visibly it measures its limitations and how carefully the team tests what may remain.

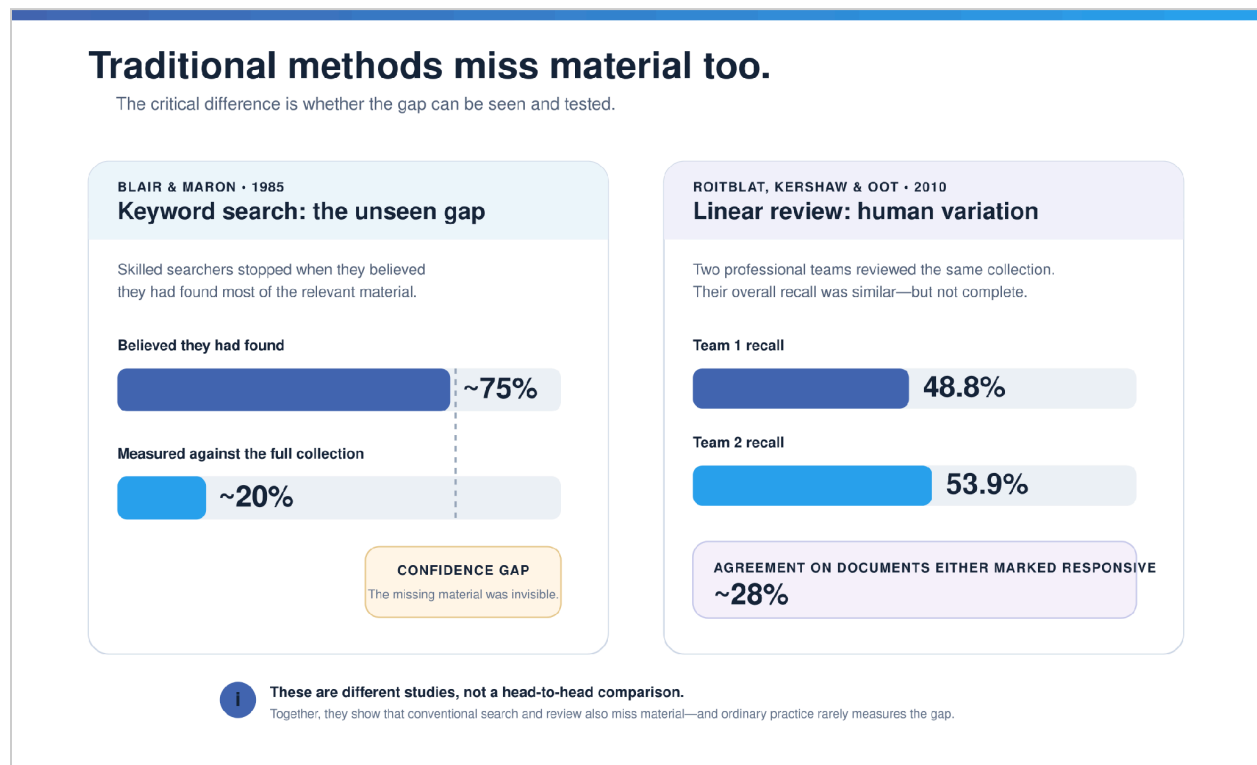


FIGURE 2. TRADITIONAL METHODS MISS MATERIAL TOO Two foundational studies show a confidence gap in keyword search and substantial variation in human review.

How three methods become one process

Hybrid retrieval begins by running a query through lexical and semantic pathways. Each produces a ranked list based on different evidence: one on the language the documents contain, the other on the meaning they express. The system then combines those signals so that a document scoring strongly on either method can surface, while a document supported by both can rise higher.



CAL joins as review begins. The first reviewer decisions provide training examples. The classifier ranks the remaining population, reviewers assess the highest-ranked documents, and the classifier retrains as more decisions are made. By the time the team has reviewed several hundred documents, CAL is no longer looking merely for the original words or a generally related concept. It is looking for the patterns that distinguish relevant from nonrelevant material in that matter.

This is addition, not replacement. Exact-match controls still matter. Semantic similarity does not excuse weak query design. CAL does not cure inconsistent review judgments. And more signals do not automatically produce a better result. Poorly designed fusion can bury an exact match, while poor training examples can distort a classifier.

A sound system does more than place three search labels on the screen. It preserves exact-match controls, provides visibility into why documents surfaced and supports validation of the unreviewed remainder.

From gate to lens

The most important change may be procedural rather than technical.

Traditional culling makes scope a gate. Documents matching the agreed terms enter the review population. Documents outside those terms do not. If the team later discovers that the vocabulary was incomplete, recovering what was excluded may require new searches, new negotiations and a revised review population.

Hybrid retrieval allows scope to operate more like a lens. The team can focus on a topic, work through the strongest material and then return to the full collection to test what the narrower view missed. Narrowing becomes a reversible decision rather than a permanent boundary.

Ranking changes the work as well. A filter divides a collection into two groups. A ranking orders it. The documents most likely to matter arrive first, and the team can observe how the value of the results changes as review proceeds.

This makes marginal return visible. At the beginning of a well-ranked review, relevant documents and new information may appear frequently. Later, the team may encounter more cumulative material: another copy of the same attachment, another participant repeating an established point, another document confirming a fact already supported several times.

A flattening curve is not proof that nothing relevant remains. It is a signal to test further. The team can review beyond the apparent plateau, approach the collection with another retrieval method and examine a blind sample from the unreviewed remainder. If those tests also produce little or no new relevant material, the decision to stop rests on evidence rather than intuition.



FIGURE 3. THE MARGINAL RETURN BECOMES VISIBLE A flattening curve is a signal to extend the review, test another angle or validate the remainder—not proof that nothing remains.

Documents and information are not the same unit

Search effectiveness is commonly measured through recall: the proportion of all relevant documents that the process identifies. For a production, recall remains a central measure because the obligation runs to responsive documents.

An investigation, case assessment or strategy exercise presents an additional question. The team needs the facts, themes, sequence of events and points of disagreement, together with the documents that support or contradict them. Document recall alone does not measure whether the team has developed that understanding.

One fact might appear in 40 documents. Find six and the team may have the fact, multiple sources and enough evidence to question a witness. The document recall for that fact is only 15 percent, but the informational value may already be substantial. The reverse is also possible: a process can retrieve a high percentage of the relevant documents while missing the one email that reveals a previously unknown fact.

Roitblat's topic-modeling research illustrates the distinction. In one study, a classifier retrieved approximately 80 percent of the known-relevant documents. Yet the topics identified across the missed documents were already present in the retrieved set. The process missed documents, but in that dataset it did not miss a topic found only in those documents.

That result is not a universal rule, and topic coverage is not a substitute for reasonable recall in a production. It does, however, point toward a second useful measure for investigations: whether continued review is still producing material information the team does not already have. The objective is not simply to accumulate more documents. It is to build a sufficiently tested and well-supported understanding of the matter, including evidence that challenges the emerging account.

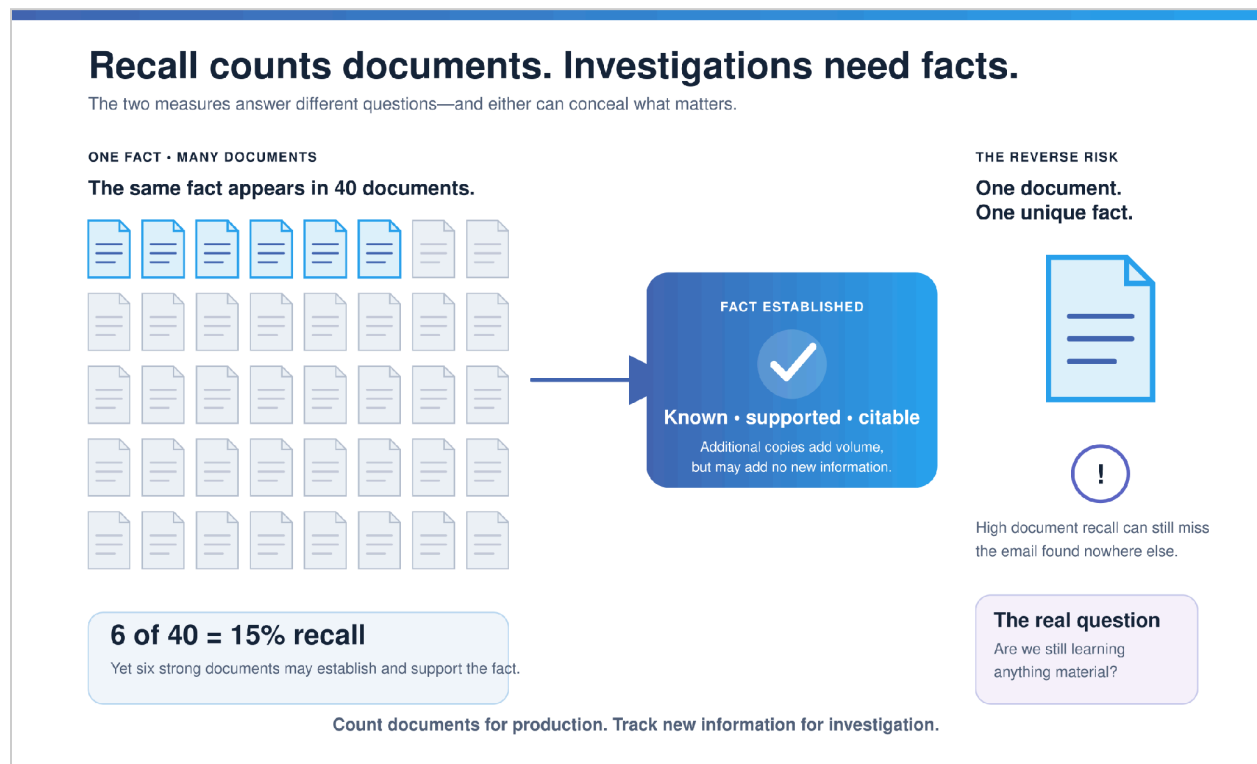


FIGURE 4. DOCUMENTS AND FACTS ARE DIFFERENT UNITS Recall counts documents. Investigations also need to track whether continued review is producing material information that is genuinely new.

A better standard for stopping

No search method promises completeness. Keyword search misses documents silently. Human reviewers disagree. Semantic retrieval can drift. CAL can inherit weaknesses in its training examples. Any claim of perfect retrieval promises something no real process delivers.

The proper comparison is not between a measurable modern process and an imaginary flawless alternative. It is between available methods, each with limitations, and the safeguards used to identify and reduce those limitations.

A ranked process offers an important advantage: the unreviewed remainder remains available for testing. The team can estimate what it still contains through sampling, compare retrieval methods, examine contrary evidence and document the basis for stopping. Measurement does not make error disappear. It makes uncertainty visible enough to manage.



This is where hybrid retrieval becomes more than a better search experience. It supports a more defensible process. The team can explain not only what it searched for, but which retrieval signals it used, how reviewer decisions shaped the ranking, what validation it performed and what the remaining population appeared to contain.

The end of one-method search

Keyword search remains indispensable. But indispensability is not sufficiency.

Legal teams no longer have to bet an investigation or production on one retrieval method and its particular blind spots. They can search for known words, search for related meaning and let the collection teach the system what relevance looks like.

The advantage is not that any method becomes infallible. It is that the methods can catch one another's mistakes - and that the team can test what remains before deciding it has searched enough.

The era of one-method search is ending. The question is whether legal teams will continue working as though nothing has changed.

References

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM*, 28(3), 289-299.

Roitblat, H. L., Kershaw, A., & Oot, P. (2010). Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1), 70-80. <https://doi.org/10.1002/asi.21233>

Roitblat, H. L. (2020). Is there something I'm missing? Topic modeling in eDiscovery. *arXiv*. <https://doi.org/10.48550/arXiv.2007.15731>

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5), 697-716. [https://doi.org/10.1016/S0306-4573\(00\)00010-8](https://doi.org/10.1016/S0306-4573(00)00010-8)



About the Authors

John Tredennick

CEO & FOUNDER, MERLIN SEARCH TECHNOLOGIES

John Tredennick is CEO and founder of Merlin Search Technologies, a company pioneering AI-powered document intelligence for legal professionals. He spent the first 20 years of his career as a trial lawyer and senior litigation partner at Holland & Hart LLP, where he became one of the first CTOs of a major law firm. In 2000, he founded Catalyst Repository Systems, an international ediscovery technology company acquired by OpenText in 2019. He has authored or edited eight books and dozens of articles on legal technology and AI, has spoken on five continents, and served as Chair of the ABA's Law Practice Management Section. He has been recognized by The American Lawyer as one of the six top ediscovery pioneers.

jt@merlin.tech



Dr. William Webber

CHIEF DATA SCIENTIST, MERLIN SEARCH TECHNOLOGIES

Dr. William Webber is Chief Data Scientist of Merlin Search Technologies. He completed his PhD in measurement in information retrieval evaluation at the University of Melbourne under Professors Alistair Moffat and Justin Zobel, and his post-doctoral research at the E-Discovery Lab of the University of Maryland under Professor Doug Oard. With more than 30 peer-reviewed publications in information retrieval, statistical evaluation and machine learning, he is a leading authority on AI and statistical measurement for information retrieval and ediscovery. At Merlin he leads the data science behind the Alchemy™ platform, including its multi-LLM architecture and the validation methods that keep AI-generated answers sourced and measurable.

wwebber@merlin.tech

