



SEARCH FINDS THE INITIAL EVIDENCE. CAL LEARNS WHAT COMES NEXT.

New research shows how relevance feedback improves what lawyers see after the initial search.

John Tredennick, J.D. and Dr. William Webber

Initially published by EDM and JD Supra in September 2026.

This is the third article in our AI Search series and its role in AI-Integrated Discovery. The full series [can be found here](#).

Search Finds The Initial Evidence. Cal Learns What Comes Next.

By John Tredennick, J.D. and Dr. William Webber

Hybrid Search finds the first 50 results. The lawyer could read all the results. But we can save the lawyer's time. An LLM can read the results, and decide which are relevant, in only a few seconds and for less than a few dollars. And then, almost instantaneously, the system knows something about relevance that it did not know before. The question is, should it make use of this knowledge to improve the rest of its retrieval?

Traditional search does not. It continues down the ranking calculated when the query first ran. Results 51 through 100 follow results 1 through 50 without regard to what the system learned from evaluating the initial results.

An adaptive system takes a different approach. It uses the LLM's clearest relevance assessments to train a classifier, then combines that classifier with the remaining search results. Passages scoring from 60 to 100 become positive training examples. Those scoring

from 0 to 40 become negative examples. The uncertain middle range, from 41 to 59, is excluded from classifier training. Search supplies the starting evidence. The LLM evaluates it. CAL learns what relevance looks like in the particular matter and changes the ranking as the investigation develops.

We tested these two paths on the same public email collection and the same twelve discovery-style topics used in our earlier Hybrid Search research. The result was clear. Across results 51 through 150, the adaptive workflow returned 82 more responsive passages than continued search and raised precision from 66.8% to 73.7%. The largest improvement came immediately after the first round of LLM assessments was used to train CAL.

The First Fifty Results Are Only the Beginning

The first two articles in this series examined how keyword and natural-language search perform when lawyers are trying to find the initial evidence. Exact words can be indispensable. Natural-language search can find relevant ideas expressed in vocabulary the lawyer did not anticipate. A merged ranking provides a practical hedge because the lawyer generally does not know which type of language the evidence will contain.

Those methods answer the opening question: what evidence should the system surface first? They do not answer the question that follows. Once an LLM has evaluated the initial results, should the system continue displaying a ranking that never changes, or should it learn from those assessments?

That question matters because a query is an imperfect statement of intent. A lawyer may be looking for evidence of shifting responsibility, ignored warnings, competing public and private interests, or the real subject behind an internal code name. The first responsive passages often reveal details that were unknown when the query was written. They identify the people involved, expose the vocabulary used inside the organization, and show which factual relationships make a passage important.

Every high-confidence relevance assessment therefore adds information. A static ranking leaves that information unused.

Testing Two Paths After the Initial Search

The experiment began with the merged-search method tested in the earlier work. An LLM expanded each request into likely keywords for BM25 search. The original request was also submitted to natural-language embedding search. The two rankings were combined using reciprocal rank fusion, with slightly greater weight given to natural-language search.

Both experimental paths received exactly the same first 50 merged-search results. The LLM evaluated each passage using the same review protocol applied in the earlier experiments. Across the twelve primary topics, 85.3% of those results scored 60 or higher and were therefore

counted as responsive. That common starting point allowed us to ask what should happen next without changing the evidence each system had initially presented.

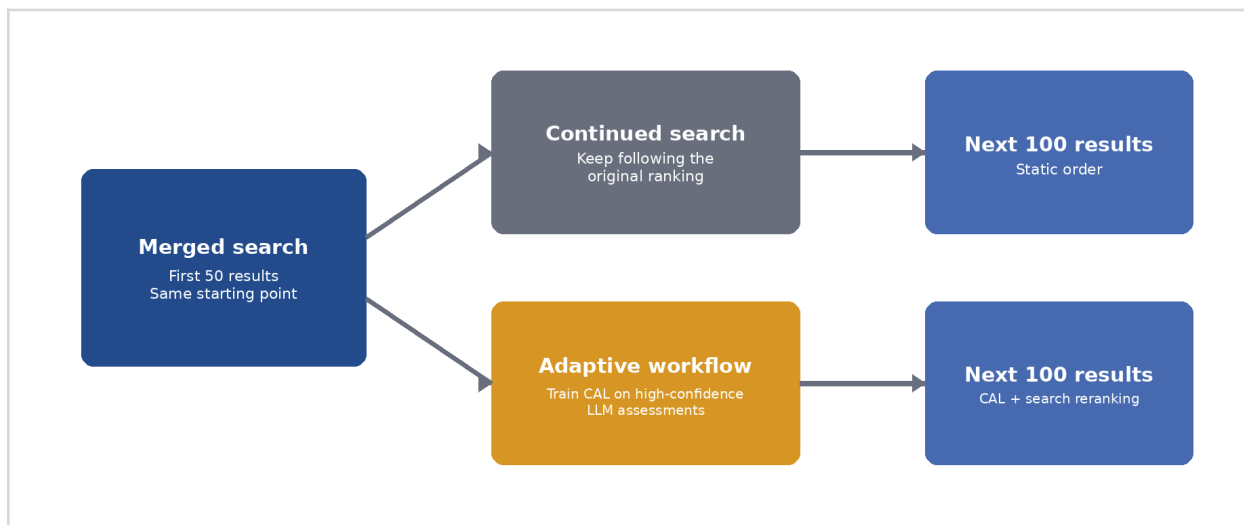


Figure 1. Both paths began with the same 50 merged-search results. One continued the original ranking. The other trained CAL on the LLM's high-confidence relevance assessments and merged CAL with the remaining search results.

The continued-search path simply moved farther down the existing merged ranking. The adaptive path trained a logistic-regression classifier on the initial LLM assessments, using scores of 60 to 100 as positive examples and scores of 0 to 40 as negative examples. Scores of 41 to 59 were excluded because they did not provide a sufficiently clear training signal. The system then combined the classifier ranking with the remaining search ranking. After the next 50 passages were assessed, the classifier was retrained using only the LLM assessments generated for passages actually presented along the adaptive path.

This was not a contest between search and a classifier operating alone. Search continued to contribute candidates throughout the experiment. The question was whether the system benefited from adding a model trained on the LLM's high-confidence assessments of what had proved responsive and nonresponsive in the matter.

LLM Assessments Improved What the System Found Next

The first adaptive batch produced the strongest individual result. Among results 51 through 100, continued merged search achieved 66.3% precision. The adaptive classifier-and-search workflow achieved 77.3%, an improvement of 11 percentage points. Put another way, the classifier-and-search approach found 17% more responsive documents in the second batch than merged search alone.

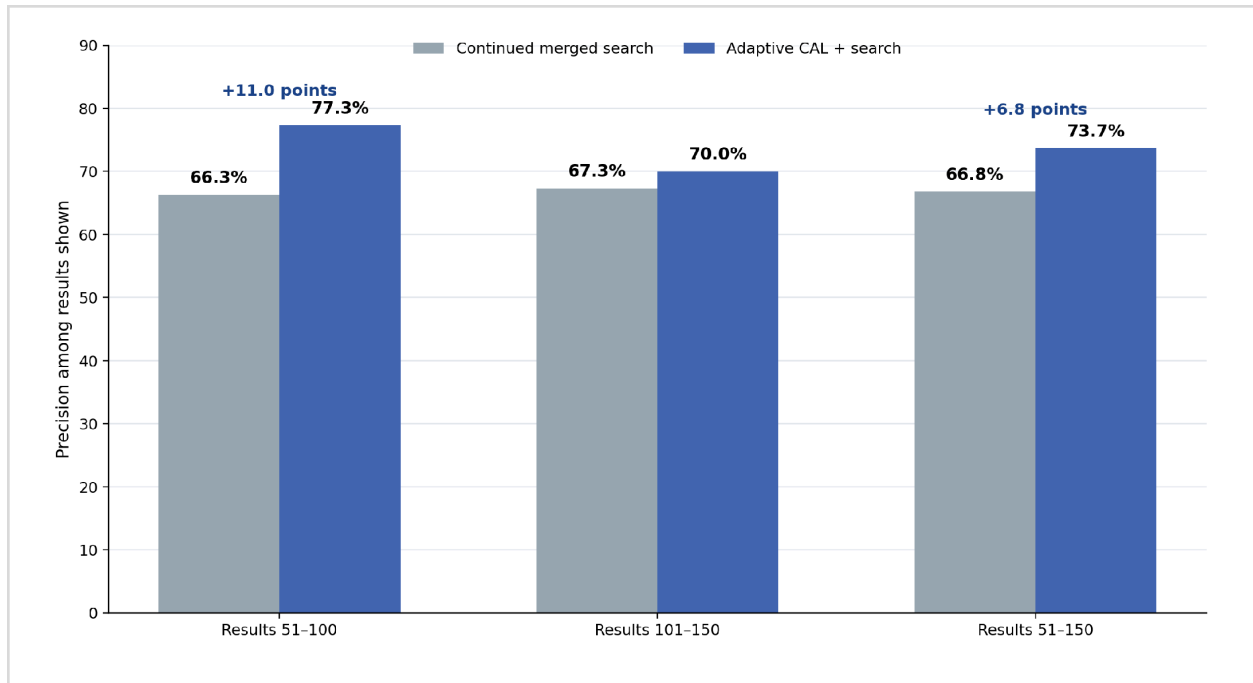


Figure 2. Adaptive CAL plus search improved precision in the next 50 results and retained a cumulative advantage across results 51 through 150.

The third batch seems less decisive when viewed by itself. For results 101 through 150, continued search achieved 67.3% precision and the enhanced adaptive workflow achieved 70.0%. The adaptive workflow won by only 2.7%. Looking at the third batch in isolation, however, understates the adaptive method's performance. Because the adaptive solution found more responsive passages in the second batch, it left itself with fewer to find in the third. The adaptive classifier-search combination was outperforming at a task that it had made harder for itself.

The cumulative comparison is fairer. Across results 51 through 150, continued search returned 802 responsive passages. The adaptive trajectory returned 884. That increased precision from 66.8% to 73.7%, a 6.8-point advantage.

Stage	Continued search	Adaptive CAL + search	Difference
Shared results 1–50	85.3%	85.3%	Same
Results 51–100	66.3%	77.3%	+11.0
Results 101–150	67.3%	70.0%	+2.7
Cumulative results 51–150	66.8%	73.7%	+6.8
Cumulative results 1–150	73.0%	77.6%	+4.6

Precision is the percentage of presented chunks receiving an LLM relevance score of 60 or higher. Scores from 41 to 59 counted as nonresponsive for this precision calculation but were excluded from CAL training.

Conceptual Topics Benefited Most at First

The first adaptive round helped both types of topics, but the largest improvement appeared on conceptual requests. For topics with distinctive terminology, adaptive precision increased by 7.3 points in results 51 through 100. For topics based on relationships, policy trade-offs, or factual patterns, it increased by 14.7 points.

That pattern makes sense. Keyword search already performs well when relevance is anchored by a distinctive name or phrase. Conceptual relevance is harder to express in advance because responsive passages may use varied language. Once the LLM identifies several high-confidence positive and negative examples, the classifier can use their embedding representations to find additional passages with similar characteristics even when they do not repeat the query's vocabulary.

The topic-level results also caution against treating adaptive learning as an automatic improvement. The largest cumulative gain came from the topic concerning public support for private sports projects, where the adaptive path returned 41 more responsive passages. Climate change gained 17, slot machines gained 11, and the balance between self-defense rights and public safety gained 10. Environmental trade-offs lost nine responsive passages, while several other topics had smaller losses. CAL improved average yield, not every individual ranking.

Project Checkmate Shows How the Methods Build on Each Other

Project Checkmate remains the clearest illustration of why hybrid retrieval needs different methods. The phrase referred to a collaboration between IBM and The Scripps Research Institute involving advanced pandemic-virus research. Nothing about the words "Project Checkmate" described that subject.

In the earlier experiment, raw keyword search found 34 responsive passages among its first 50 results. Natural-language search found none. Merged search returned 17. The literal phrase gave keyword search an immediate advantage, while semantic similarity had no meaningful signal from the code name alone.



Figure 3. Keyword search found the opaque code name. The LLM's relevance assessments then supplied examples that allowed CAL to learn the subject behind it.

Those 17 responsive merged-search passages, together with the high-confidence negative examples among the initial results, were nevertheless enough to train a classifier. Among results 51 through 150, continued merged search returned 17 additional responsive passages. The adaptive workflow returned 24, or 40% more. Across the first 150 results, the adaptive path found 41 responsive passages, seven more than raw keyword search found across its own first 150. Raw keyword search found all 34 of its responsive passages in the first 50 and none in either later batch. Adaptive search found relevant documents much faster than static merged search, and found relevant documents that raw keyword search couldn't.

The lesson is not that CAL made the code name unnecessary. The literal match supplied the seed evidence. The LLM's assessments then supplied positive and negative examples from which the classifier could learn what the code name represented in this collection. Keywords found the name. CAL learned the subject behind it.

CAL Adds a Third Kind of Intelligence

Keyword and natural-language search derive their rankings from the query. CAL derives its ranking from relevance judgments. In the workflow tested here, an LLM generated those judgments automatically by evaluating the initial results. That difference gives each method a separate role.

- Keyword search finds exact names, identifiers, phrases, citations and other distinctive language.
- Natural-language search finds passages that express related meaning even when their vocabulary differs from the request.
- Continuous active learning builds a matter-specific model from high-confidence positive and negative examples and repeatedly reprioritizes what remains.

The tested workflow combined all three. Initial keyword and natural-language search produced a strong first group of results. The LLM then scored each passage for relevance. CAL learned

from those assessments. Search continued to contribute candidates while the classifier increasingly reflected what had proved responsive in the matter.

This is the practical value of Hybrid Search. The methods do not merely occupy three tabs in an interface. They transfer information from one stage to the next. Search finds examples. The LLM evaluates them. High-confidence assessments train CAL. CAL improves the next ranking.

What the Study Establishes

The experiment supports a focused conclusion. Using LLM assessments to train CAL improved the yield of the results presented after the initial search. The largest gain appeared in the first adaptive batch, and the workflow retained a cumulative advantage through 150 results. The evidence also supports combining classification with continued search rather than treating CAL as a replacement for retrieval.

The study does not establish that classification improved every topic or every batch. It does not compare a classifier operating alone with search. It does not show that every collection will produce the same effect. The relevance judgments were generated by an LLM on individual chunks, and the same model family was used for query expansion and assessment. The experiment used one public email collection and twelve deliberately constructed topics.

Most important, this was not a collection-wide recall study. Precision measures the proportion of presented results that were responsive. It does not reveal how many responsive passages remained elsewhere in the collection. CAL is designed to pursue relevant material beyond the initial search, but these results establish better ranking and responsive yield within a fixed review budget, not exhaustive recall.

Research Methodology

Element	Research design
Collection	Public Jeb Bush email collection, evaluated at the document-chunk level.
Topics	Twelve primary topics: six with distinctive terminology and six based on conceptual relationships or policy trade-offs. Project Checkmate was retained as a supplementary code-name topic.
Initial search	LLM-expanded BM25 keyword search and natural-language embedding search combined through weighted reciprocal rank fusion.
Adaptive model	Logistic regression trained on stored embedding vectors. Scores from 60 to 100 supplied positive examples; scores from 0 to 40 supplied negative examples. Scores from 41 to 59 were excluded from training.
Comparison	Continue the merged-search ranking or train a classifier on the LLM assessments and merge its ranking with the remaining search results.
Assessment	One-chunk relevance assessments generated by the same LLM used in the earlier experiments. Confidence intervals resampled topics rather than treating chunks as independent.

Search Should Learn as the Investigation Develops

A lawyer's first question rarely captures everything that later proves important. The early evidence changes the team's understanding of the people, terminology and factual relationships involved. Hybrid Search finds that evidence, and an LLM evaluates it. Those assessments generate information the system can use immediately to improve retrieval.

Continuing down a static ranking leaves those assessments unused. An adaptive workflow treats them as evidence. Our results show that using the LLM's high-confidence assessments improved what the system found next, especially in the first batch after training, and produced a meaningful cumulative advantage across the next 100 results.

Hybrid Search therefore has a natural progression. The lawyer defines the question. Keywords find exact language. Natural-language search finds related meaning. The LLM evaluates the initial evidence. CAL learns from the clearest positive and negative assessments and helps decide what should come next.

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.

webber@merlin.tech

