
# Quantitative Performance Benchmarks for Keyword Search and Technology-Assisted Review in eDiscovery and Digital Forensics

*Research*

Fookes Software Ltd
Charmey, Switzerland
www.aid4mail.com

# Table of Contents

# Quantitative Performance Benchmarks for Keyword Search and Technology-Assisted Review in eDiscovery and Digital Forensics

**Date:** 2026-03-31
**Scope:** Defensible quantitative baselines for keyword-based search and Technology-Assisted Review (TAR), for use in comparative evaluation of document classification systems

## 1. Executive Summary

This document synthesizes the best available quantitative evidence on the performance of two principal document classification methodologies used in eDiscovery and digital forensics: keyword-based search and Technology-Assisted Review (TAR). All performance figures are derived from peer-reviewed studies, standardized competitive evaluation tracks (TREC Legal Track and TREC Total Recall Track), and documented industry evaluations. Only findings with explicit quantitative reporting have been included.

### Keyword Search — Benchmark Range Summary

| Metric | Typical Range | Best Observed (Benchmark Conditions) | Evidence Quality |
|---|---|---|---|
| Recall | 20%–40% | ~76% (engineered deterministic query, single topic) | Well-established |
| Precision | 10%–79% (query-dependent) | ~84% (same engineered query) | Well-established |
| F1 | ~25%–40% | ~80% (same engineered query) | Well-established |
| Miss rate | 60%–80% | ~24% (best case) | Well-established |

**Interpretation:** Keyword and Boolean search methods consistently underperform practitioner expectations in recall. The foundational Blair & Maron (1985) study found that experienced legal professionals, after iterative refinement, achieved only 20% average recall while believing they had reached ≥75%. This perception gap is reproducible across decades of subsequent testing. Reference Boolean queries in TREC Legal Track evaluations ranged from under 4% to approximately 24% recall across different evaluation years and judgment regimes. Only a rigorously engineered, iteratively measured deterministic query workflow achieved recall above 75%, and that result applies to a single benchmark topic under controlled conditions.

### TAR — Benchmark Range Summary

| Method | Typical Recall | Typical Precision | Typical F1 | Evidence Quality |
|---|---|---|---|---|
| TAR 1.0 (SPL/SAL) | 50%–75% | 60%–80% | 55%–75% | Moderately supported |
| TAR 2.0 (CAL) | 75%–96% | 80%–96% (at moderate-to-high prevalence; see §4.5) | 75%–96% | Well-established |
| Hybrid CAL + attorney judgment | 82%–100% (F1) | High | 82%–100% | Moderately supported |
| Human linear review (baseline) | 49%–54% | 18%–20% | ~27%–28% | Moderately supported |

**Interpretation:** CAL-based TAR (TAR 2.0) is the best-evidenced method for achieving high recall at competitive precision. Multiple controlled evaluations on large email collections show recall of 90%–96% and precision of 80%–96% simultaneously. TAR 1.0 (passive or simple active learning) achieves recall comparable to human linear review at better precision. Human linear review is not the reliable baseline it has historically been assumed to be.

### Cross-Cutting Caveats

- The **65% inter-assessor agreement limit** (Voorhees, 2000) constrains evaluation: because human experts reviewing the same corpus agree on relevance only ~65% of the time, measured agreement between any automated method and a single human gold standard cannot be reliably interpreted above this threshold. Systems may correctly classify documents that the reference assessor missed, but such correct classifications will appear as errors within the evaluation framework.
- **Prevalence (dataset richness)** is a primary determinant of precision. In low-prevalence collections (<5% relevant), even a classifier with high recall and low false-positive rate may produce poor observed precision due to the mathematics of base rates.
- **Digital forensics** lacks a TREC-equivalent benchmark. eDiscovery benchmarks are the best available proxy; forensics-specific evidence is limited to post-retrieval ranking studies.

## 2. Scope and Definitions

### 2.1 Methodologies Covered

#### Keyword-Based Search

Any method that identifies documents by matching document text against one or more explicitly specified terms or expressions. Three sub-variants are distinguished:

- **Simple keyword search / string search:** Direct matching of one or more terms against indexed text. Used in eDiscovery for culling and in digital forensics for byte-level searching across disk images.
- **Boolean search:** Formal logical queries combining terms with AND, OR, NOT, and proximity operators. The dominant form of keyword search in legal eDiscovery platforms.
- **Iterative / negotiated keyword refinement:** A workflow in which Boolean or keyword queries are repeatedly revised based on interim results and attorney feedback, including the formally negotiated queries used in TREC Legal Track reference runs.

All three rely on exact lexical matching and are fundamentally subject to vocabulary mismatch (synonymy and polysemy).

### Technology-Assisted Review (TAR)

The application of supervised machine learning to document classification, using human-coded training examples to extrapolate relevance decisions across a larger unreviewed corpus. Three protocol variants are distinguished:

- **TAR 1.0 / Simple Passive Learning (SPL):** A static seed set is used to train the model, which is then applied to the full corpus in a single pass. Training quality depends entirely on the initial seed set; the model is not updated after deployment.
- **TAR 1.0 / Simple Active Learning (SAL):** The classifier iteratively selects documents for human review based on uncertainty or predicted relevance, then retrains. Stops when the model stabilizes.
- **TAR 2.0 / Continuous Active Learning (CAL):** The classifier is updated continuously throughout the entire review workflow. Each human coding decision refines the model in real time, eliminating the need for a formal stopping criterion tied to model stability.

These three protocols yield materially different performance outcomes and must not be conflated when interpreting benchmarks.

## 2.2 Evaluation Metrics

| Metric | Definition | eDiscovery / Forensics Implication |
|---|---|---|
| **Recall** (Sensitivity; TPR) | TP / (TP + FN) | Fraction of truly relevant documents retrieved; inverse of the false negative (miss) rate |
| **Precision** (PPV) | TP / (TP + FP) | Fraction of retrieved documents that are truly relevant; inverse of the false positive (over-collection) rate |
| **False Positive Rate (FPR)** | FP / (FP + TN) | Rate at which non-relevant documents are incorrectly flagged |
| **False Negative Rate (FNR; Miss Rate)** | FN / (FN + TP) | 1 − Recall; rate at which relevant documents are missed |

| | | |
|---|---|---|
| **F1 Score** | 2 × (Precision × Recall) / (Precision + Recall) | Harmonic mean balancing recall and precision; penalizes extreme imbalances |
| **Elusion** | Relevant documents in predicted-non-relevant set / total predicted-non-relevant | Quality control metric; measures the miss rate within the discarded pile |

F1 is used throughout as a combined effectiveness metric, but recall is the primary metric of concern in eDiscovery and forensic review workflows, where false negatives carry greater legal and investigative risk than false positives.

### 2.3 Study Inclusion Criteria

Only sources satisfying all of the following criteria have been used to establish benchmark ranges:

1. Report at least one of: recall, precision, FPR/FNR, or F1 as an explicit numeric value.
2. Apply to a document corpus of meaningful scale (not toy datasets).
3. Use a defined and repeatable evaluation methodology.
4. Are either peer-reviewed or produced by a recognized standardized evaluation track (TREC).

Industry practitioner reports are included where the dataset and protocol are disclosed and the findings are internally consistent with peer-reviewed evidence. Findings supported by a single non-peer-reviewed source are identified as such and not used to establish typical ranges.

## 3. Keyword Search Performance Benchmarks

### 3.1 Primary Evidence Table

| Source | Year | Dataset | Method | Recall | Precision | F1 | Notes |
|---|---|---|---|---|---|---|---|
| Blair & Maron, *Comm. ACM* | 1985 | ~40,000 legal/corporate documents (BART litigation) | Keyword/Boolean (STAIRS); iterative refinement | **20.0% avg** (range 2.8%–78.7%) | **79.0% avg** (range 19.6%–100%) | ~32% at avg operating point | Foundational field study. Attorneys estimated 75% recall; actual was 20%. High precision reflects over-specificity of queries, not system capability. |
| TREC 2007 Legal Track, reference Boolean run (refL07B) | 2007 | IIT CDIP / Enron corpus (~7 million documents) | Negotiated Boolean query (reference run) | **22% est.** | **29% est.** | ~25% | Reference Boolean run negotiated between opposing |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | counsel. Mean recall of 22% means the final negotiated query missed ~78% of relevant documents on average across all 43 Ad Hoc topics. Per-topic recall ranged from 0% to 100%. |
| TREC 2008 Legal Track, reference Boolean run | 2008 | IIT CDIP tobacco litigation corpus | Negotiated Boolean query (refL08B) | **24.0% est.** | **28.0% est.** | ~25.7% | Estimated from sampled judgments. Scores reflect the difficulty of maintaining both metrics simultaneously. |
| TREC 2009 Legal Track, reference Boolean query | 2009 | Enron FERC email corpus (~520,000 messages) | Final negotiated Boolean query | **<4% avg** | **~39% avg** | **~0.06 avg** | Under test judgments. The same query achieved ~17% recall under training judgments, illustrating severe sensitivity to judgment regime. |
| Grossman & Cormack (JOLT Study), H5 Topic 204 | 2011 (task year 2009) | TREC 2009 Legal Track Interactive Task | Engineered deterministic keyword query (non-ML) | **76.2%** | **84.4%** | **80.1%** | Single completed topic. Iteratively engineered, explicitly measured workflow. Not representative of typical keyword search; represents best-case with exceptional protocol discipline. |
| Baron (NARA / Sedona Conference) | 2004–2007 | U.S. government email archives (large-scale) | Boolean keyword search | **~22%** | Not reported | — | Non-peer-reviewed; cited in secondary literature but primary source |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | | | | | | | not independently verified. Convergent with Blair & Maron in large-volume email context. |
| Tredennick & Bye (Catalyst Repository) | ~2017 | Real litigation matter (undisclosed) | Keyword-only culling | **~39%** | Not reported | — | Single-matter industry case study; primary source not independently verified. Recall figure is at the upper end of practitioner-reported iterative keyword results. Not generalizable. |
| Dimm (Clustify platform), competitive benchmark | 2018 | Law, medical, and biological datasets | Iterative keyword search by motivated human participants | **5.5%–38.1%** (by dataset) | Not reported | — | Industry study; not peer-reviewed. Recall ranges are consistent with academic evidence. TAR outperformed keyword on every dataset in the same evaluation. |

*Note: TREC entries specify the evaluation track year and run identifier. Where a reference run from a prior year was carried forward, both the originating year and evaluation year are noted. All TREC metrics are cross-topic means unless stated otherwise.*

## 3.2 Aggregated Performance Ranges

### Recall

- **Baseline (single-pass keyword search):** 20–25%, anchored by Blair & Maron and corroborated by NARA data.
- **Iterative and negotiated keyword search (typical production use):** 20–40%. Even with extended refinement cycles, published evidence rarely supports claims above 40%.
- **Reference negotiated Boolean queries (TREC Legal Track):** 4%–24%, with high variability depending on judgment methodology and topic set.
- **Best case (rigorously engineered, explicitly measured, benchmark conditions):** ~76%, documented for a single topic under exceptional protocol discipline. This figure is not generalizable to typical production workflows.
- **Typical miss rate:** 60%–80% at baseline; 55%–65% with iterative refinement.

### Precision

- **High-precision / low-recall operating point (narrow queries):** 70%–79%. Achievable when queries are highly targeted and recall is not a priority.
- **Moderate-recall / low-precision operating point (broad queries):** 10%–30%. When Boolean queries are broadened to increase recall, precision collapses substantially. The TREC 2007 reference run achieving 22% recall at 29% precision illustrates this trade-off, where even a professionally negotiated Boolean query missed ~78% of relevant documents.
- **Typical range across observed operating points:** 10%–79%, depending entirely on query specificity.

### F1 Score

- **At typical operating points in production keyword search:** approximately 25%–40%.
- At the Blair & Maron average operating point (20% recall, 79% precision): F1 ≈ 32%.
- At the TREC 2007 reference Boolean operating point (22% recall, 29% precision): F1 ≈ 25%.

## 3.3 Key Conditions Affecting Keyword Performance

**Vocabulary mismatch** is the fundamental failure mode. Document authors and legal searchers use different terminology to describe the same events and concepts. In the Blair & Maron study, attorneys searched for “disaster” and “accident” while relevant documents used phrases such as “the unfortunate situation” to avoid liability framing. Exact string matching cannot bridge semantic equivalences, causing systematic false negatives regardless of query iteration effort.

**Query specificity and the precision–recall trade-off** are inseparable in keyword search. Narrowing Boolean operators and phrase requirements improves precision but reduces recall; broadening queries increases recall at the cost of precision. No keyword architecture resolves this trade-off without introducing a different classification layer.

**Collection characteristics** govern keyword effectiveness. Performance is better on narrow topical collections with consistent vocabulary (regulatory filings, named entities, product codes) and worse on informal communications (email, messaging), large multi-custodian collections, and queries requiring conceptual or behavioral criteria (e.g., “awareness of risk”).

**Iterative refinement** extends recall from the ~20% baseline toward ~35%–40%, but evidence for keyword recall exceeding 50% through query iteration alone is sparse and methodologically weak in the published literature.

**Judgment instability** affects measured recall substantially. The same TREC 2009 reference Boolean query produced recall estimates of ~17% under training judgments and under 4% under test judgments—a five-fold difference from the same query on the same corpus under different assessor regimes.

### 3.4 Evidence Quality Assessment

| Claim | Quality |
|---|---|
| Keyword recall typically 20%–25% without iteration | **Well-established** (Blair & Maron 1985; NARA; TREC Legal Track) |
| Keyword recall can reach 35%–40% with iterative refinement | **Moderately supported** (industry practitioner data; limited controlled studies) |
| Keyword precision 70%–79% is achievable at low-recall operating points | **Well-established** (Blair & Maron 1985) |
| Precision collapses substantially as recall increases | **Well-established** (information retrieval theory; TREC Legal Track empirical data) |
| Recall >50% from keyword methods alone is not reliably achievable | **Moderately supported** (consistent across all sources; limited controlled data at higher recall levels) |

## 4. TAR Performance Benchmarks

### 4.1 Primary Evidence Table

| Source | Year | Dataset | Method | Recall | Precision | F1 |
|---|---|---|---|---|---|---|
| TREC 2009 Legal Track Interactive Task (Waterloo/Cormack) | 2009 | Enron FERC email corpus (~520,000 messages) | TAR: human-in-the-loop active learning | **67.3%–86.5%** (across 4 topics) | **69.2%–91.2%** | **76.4%–84.0%** |
| TREC 2009 Legal Track Interactive Task (H5) | 2009 | Same corpus | TAR: active learning + iterative search | **≥70%** (multiple runs) | **≥70%** (simultaneously) | **≥0.70** (6 runs) |
| Roitblat, Kershaw & Oot, *JASIST* — human review teams | 2010 | Real eDiscovery matter (~2.3M documents; DOJ second request; majority emails) | Human linear review (professional reviewers; two independent teams) | **49% (Team A), 54% (Team B)** | **20% (A), 18% (B)** | **0.27 (A), 0.28 (B)** |
| Roitblat, Kershaw & Oot, *JASIST* — automated systems | 2010 | Same corpus | TAR automated classification (two commercial systems) | **45.8%–52.7%** | **27.1%–29.5%** | **0.34–0.38** |

| | | | | | | |
|---|---|---|---|---|---|---|
| Grossman & Cormack (JOLT Study) | 2011 | TREC 2008/2009 tasks + real legal matters | TAR: active learning vs. exhaustive manual review | Comparable to or exceeding manual review | Higher than manual review | Higher than manual |
| Cormack & Grossman, *SIGIR* | 2014 | 4 TREC 2009 topics + 4 real matters | CAL vs. SAL vs. SPL (head-to-head, same data) | CAL highest at all recall levels | CAL highest | CAL highest |
| Grossman & Cormack, *SIGIR* ("Roger and Me") | 2017 | Kaine gubernatorial email corpus (401,960 emails, 3 topics) | CAL with adjudication strategy vs. expert manual review | **CAL: 90%–96%** (per topic); avg 93% | **CAL: 80%–96%** (per topic); avg 88% | **CAL: 0.87–0.96**; avg 0.91 |
| Losey / e-Discovery Team, TREC Total Recall Track | 2015–2016 | Jeb Bush email corpus (290,099 emails; 30–34 topics) | Hybrid CAL + attorney judgment | **F1 82%–100%** across topics | High (concurrent with F1) | **0.82–1.00** |

## 4.2 Protocol Hierarchy: SPL, SAL, and CAL

The Cormack & Grossman (SIGIR 2014) study is the most rigorous head-to-head comparison of TAR protocols on identical data. Its findings establish an unambiguous performance hierarchy:

| Protocol | Typical Recall | Typical Precision | Evidence Quality |
|---|---|---|---|
| SPL (Simple Passive Learning) | 50%–65% | 60%–75% | Moderate |
| SAL (Simple Active Learning) | 65%–77% | 65%–80% | Moderate |
| **CAL (Continuous Active Learning)** | **75%–96%** | **80%–96%** | **Well-established** |

CAL eliminates the static training-phase problem inherent in SPL and SAL. Because the model is updated with every human coding decision throughout the review, it adapts to evolving relevance concepts and is less sensitive to the quality or representativeness of the initial seed set.

## 4.3 Aggregated Performance Ranges

### Recall

- **TAR 1.0 (SPL/SAL) — typical range:** 50%–75%.
- **TAR 2.0 (CAL) — typical range:** 75%–96%, well-documented across multiple controlled evaluations.
- **Industry / legal defensibility threshold:** Recall levels around 75% are commonly discussed in U.S. federal discovery practice as reasonable targets for defensible

TAR workflows (see, e.g., Grossman & Cormack Glossary, 2013; *Hyles v. New York City*, S.D.N.Y. 2016).

- **Best observed (controlled benchmark):** 96% recall at 96% precision (Kaine email corpus, Legal Hold topic).

### Precision

- **Human linear review baseline:** 18%–20% (Roitblat et al. 2010).
- **TAR 1.0 at typical operating points:** 60%–80%.
- **CAL at typical operating points:** 80%–96%.

### F1 Score

- **TAR 1.0 (TREC 2009 Interactive Task, top runs):** 0.70–0.84 across 24 submitted runs.
- **CAL / TAR 2.0 (Kaine email study, per-topic):** 0.87–0.96.
- **Hybrid CAL + attorney (TREC Total Recall Track):** 0.82–1.00 across 30+ topics.

### Review Effort Reduction

Effort reduction is a secondary but operationally significant outcome. In the Kaine email study (Cormack & Grossman, 2017), the CAL strategy required review of 136,993 documents—a 56% reduction from the 310,196 documents reviewed in the manual baseline—at slightly superior accuracy. In the TREC 2009 Interactive Task, the Waterloo active learning workflow reviewed 0.5%–4.1% of the full collection per topic while achieving recall of 67%–87%. Industry synthesis across multiple matters suggests that TAR workflows typically reduce total review volume by 50%–80% compared to exhaustive manual review or broad keyword culling, though this range is not derived from a single controlled study.

## 4.4 The Roitblat et al. (2010) Ground-Truth Caveat

The Roitblat et al. study is frequently cited to support TAR effectiveness, but its measurement context requires careful interpretation. The study measured automated systems and human re-review teams against an original production by 225 attorneys—a baseline that was itself not independently adjudicated. The authors explicitly acknowledged that the metrics “imply a stable ground truth” that they did not claim to have. As a result, the nominal recall figures (45.8%–52.7%) and precision figures (27.1%–29.5%) for automated systems in this study likely understate true performance relative to an independently established gold standard. The study’s primary contribution is its demonstration that automated systems achieve F1 scores at least as high as—and in practice higher than—professional human reviewers operating against the same imperfect baseline.

## 4.5 Conditions Affecting TAR Outcomes

**Prevalence (richness)** exerts a strong influence on precision at any given recall level. In low-prevalence collections (<1%–2% relevant), the classifier must achieve very high specificity to prevent false positives from overwhelming true positives in absolute numbers. CAL is better suited to low-prevalence environments than SPL/SAL because it continues to refine the model as rare positive examples are found.

**Seed set quality** disproportionately affects TAR 1.0. A seed set drawn from a narrow keyword search or a single custodian will produce a model that overfits those features,

suppressing recall on conceptually related documents using different language. CAL mitigates this by not depending on an upfront seed set.

**SME (Subject Matter Expert) consistency** is a primary determinant of TAR quality that is often underappreciated. Grossman & Cormack (2012) demonstrated that human reviewers frequently disagree with each other on relevance, and that inconsistency in training labels degrades model performance directly. TAR quality is bounded from above by the consistency of the human feedback it receives.

**Stopping rules** affect the precision–recall trade-off in TAR 1.0 / SAL workflows. Stopping too early inflates precision but depresses recall; stopping too late wastes review effort. CAL largely bypasses this problem by continuing to classify throughout the review process.

### 4.6 Evidence Quality Assessment

| Claim | Quality |
|---|---|
| CAL achieves recall 75%–96% in controlled evaluations | **Well-established** (multiple peer-reviewed studies; TREC Legal Track; Grossman & Cormack 2014, 2017) |
| CAL precision 80%–96% at 75%–96% recall | **Well-established** (SIGIR 2017 “Roger and Me”; Kaine email study) |
| CAL statistically outperforms SAL and SPL | **Well-established** (SIGIR 2014, $p<0.01$, 8 test cases) |
| TAR 1.0 (SPL/SAL) recall 50%–75% | **Moderately supported** (TREC 2009; Roitblat et al.; fewer head-to-head comparisons against CAL) |
| Human linear review achieves 49%–54% recall with 18%–20% precision | **Moderately supported** (Roitblat et al. 2010; imperfect gold standard) |
| TAR outperforms keyword search at any given recall target | **Well-established** (converging evidence from TREC Legal Track, Grossman & Cormack, Roitblat et al.) |

## 5. Digital Forensics: Applicable Evidence

### 5.1 Context and Scope Limitation

Unlike eDiscovery, the digital forensics and incident response (DFIR) community has not produced an equivalent of the TREC Legal Track: no standardized benchmark exists that measures keyword search or TAR recall/precision on forensic collections under controlled, reproducible conditions. The evidence base is therefore narrower, and eDiscovery benchmarks should be treated as the best available proxy for general classification performance.

One forensics-specific quantitative study provides directly applicable data.

## 5.2 Keyword / String Search in Digital Forensics

Forensic disk and memory analysis tools conventionally operate with a design objective of **100% recall by construction**: all byte-level matches are returned, with no filtering. The consequence is precision that approaches zero in large, noisy disk images.

Beebe & Liu (2014) quantified this baseline on a synthetic forensic disk image (the M57 Patents corpus), finding that an unranked string search returned hits at a precision of approximately **4%**—meaning 96% of investigative hits were irrelevant noise requiring manual adjudication. This figure is not a flaw in the tool; it is a direct consequence of the design philosophy of completeness over selectivity.

## 5.3 Machine Learning Ranking of Forensic String Hits

Beebe & Liu (2014) further evaluated a Support Vector Machine (SVM) classifier trained to rank forensic string search hits by relevance, using 18 features including temporal proximity, file structure, and metadata. Results by review depth are summarized below:

| Fraction of hits reviewed | Recall (allocated model) | Precision (allocated model) |
|---|---|---|
| 25% | 0.42 | 0.84 |
| 50% | 0.80 | 0.80 |
| 75% | 0.96 | 0.64 |
| 100% (all hits) | 1.00 | 0.50 (prevalence-limited) |

Overall classifier accuracy: 81.0% for allocated files, 85.9% for unallocated clusters.

**Interpretation:** These results are best understood as benchmarks for **triage efficiency** in forensic hit review, not for end-to-end legal responsiveness classification. By reviewing the top 25% of ranked hits, an investigator can achieve recall of ~0.42 at precision of ~0.84—a four-fold reduction in review burden compared to exhaustive review, without the typical recall sacrifice of conservative filtering. Reviewing the top 50% achieves an approximately equal precision–recall balance at 0.80/0.80. This represents a meaningful improvement over unranked string search but still requires human adjudication at scale.

## 5.4 ML-Based Forensic Triage of Web Artifacts

Emerging work in ML-based forensic triage of web artifacts—using approaches such as LSTM networks and autoencoders for behavioral clustering—shows promising results for classification of user session logs and browser artifacts, but no standardized benchmarks comparable to TREC exist in this area. This work is noted as indicative of machine learning's potential in forensic triage but is not used to establish generalizable benchmarks.

### 5.5 Summary for Digital Forensics

| Method | Recall | Precision | Context |
|---|---|---|---|
| Unranked forensic string search | 100% (by design) | ~4% | Baseline; all hits returned |
| SVM-ranked hits (top 25% reviewed) | ~42% | ~84% | Triage efficiency benchmark |
| SVM-ranked hits (top 50% reviewed) | ~80% | ~80% | Triage efficiency benchmark |
| SVM-ranked hits (top 75% reviewed) | ~96% | ~64% | Triage efficiency benchmark |

No equivalent of eDiscovery TAR benchmarking (end-to-end responsiveness classification at corpus scale) has been published for digital forensics under standardized, reproducible conditions.

## 6. Comparative Observations

### 6.1 Recall Characteristics

Keyword search and TAR operate at fundamentally different recall ceilings under production conditions. The typical recall range for iterative keyword search (20%–40%) falls below the typical lower bound for CAL-based TAR (75%–96%). TAR 1.0 occupies an intermediate range (50%–75%) that is comparable to—and, at typical operating points, superior to—human linear review.

The critical implication is miss rate. At a keyword baseline recall of 20%–40%, between 60% and 80% of relevant documents are not retrieved. At CAL recall of 75%–96%, between 4% and 25% are missed. The practical significance of this gap depends on the stakes of the review and the consequences of a missed document, but the gap itself is substantial and reproducible.

### 6.2 Precision Characteristics

Precision behavior differs markedly between methods. Keyword search precision is primarily determined by query specificity and exists along an inverse continuum with recall: high precision (70%–79%) is achievable only at low recall (20%–25%), and moderate recall (38%–40%) collapses precision to 10%–20%. This inverse relationship is a structural feature of lexical matching, not a calibration problem.

TAR, and CAL in particular, largely decouples precision from recall within its operating range. CAL can maintain precision of 80%–96% while achieving recall of 75%–96%, because the model learns the semantic and contextual features of relevance rather than relying on specific term matches. Precision in TAR is primarily governed by prevalence and training consistency, not by a forced trade-off with recall.

### 6.3 Operational Implications

**For legal eDiscovery:** The practical consequence of low keyword recall is that relevant documents remain undisclosed. In adversarial legal proceedings, this creates sanctions risk (adverse inference instructions, spoliation findings). TAR shifts the compliance risk from vocabulary coverage to training quality, which is arguable but auditable. The ~75% recall level commonly discussed as a reasonable target in U.S. discovery practice is achievable by CAL but not by typical keyword workflows.

**For digital forensics investigations:** Forensic investigators who require high confidence in the completeness of their findings should not rely on unranked string search for triage of large collections. ML-based ranking of string hits demonstrates that reviewing the top 50% of ranked results can achieve approximately 80% recall and 80% precision—a substantially better operating point than either exhaustive unranked review or conservative threshold filtering alone.

**The precision–recall trade-off is unavoidable for all methods.** Pushing recall toward 90%+ at acceptable precision requires either a well-trained CAL model, a hybrid attorney-plus-CAL workflow, or an exceptional keyword engineering effort that cannot be generalized. The question for any classification system is not whether the trade-off exists, but at what recall level it can maintain acceptable precision.

### 6.4 Review Effort

TAR workflows, particularly CAL, substantially reduce the total volume of documents requiring human review while maintaining or improving recall and precision compared to exhaustive methods. Evidence from multiple studies suggests review effort reductions of 50%–80% are achievable. This is operationally significant in large eDiscovery matters but is a secondary consideration relative to accuracy in most forensic contexts.

---

## 7. Limitations and Evidence Gaps

### 7.1 Inter-Assessor Agreement as an Evaluation Constraint

Voorhees (2000), analyzing TREC Ad Hoc relevance assessments across multiple topics, established that independent human expert assessors agree on document relevance approximately **65% of the time**. This is not a weakness of any specific study; it reflects genuine ambiguity in relevance determination, especially for conceptual or behavioral relevance criteria.

This ~65% agreement rate constrains evaluation:

- Measured agreement between any automated method and a single human gold standard cannot be reliably interpreted above ~65%.
- Apparent “errors” in automated classification may partly reflect inconsistency in the gold standard itself.
- When an automated system exceeds ~65%–70% agreement with a human gold standard, it may be correctly identifying documents that the reference assessor missed.

## 7.2 Gold Standard Dependency

The Roitblat et al. (2010) study illustrates a broader problem: in many real-matter evaluations, there is no independently adjudicated ground truth. Performance is measured against prior human productions that are themselves incomplete and inconsistent. Published recall and precision figures from such studies systematically understate true system performance relative to the ideal gold standard that does not exist in practice.

## 7.3 Dataset Prevalence Effects

Prevalence is rarely controlled for across eDiscovery benchmarks. Published studies draw from datasets with widely varying richness (0.5%–40%+ relevant). At very low prevalence (<1%), even a classifier with a 1% false positive rate may produce a reviewed set that is <50% relevant, regardless of recall. Reported precision figures cannot be compared across studies without knowing the dataset prevalence, and this information is often not disclosed in practitioner reports.

## 7.4 TREC Reference Boolean Query Instability

TREC Legal Track data must be interpreted with awareness that the same negotiated Boolean query can produce recall estimates that vary five-fold depending on whether training or test judgments are applied (e.g., ~17% vs. <4% for the same TREC 2009 query). This is a function of sampling-based recall estimation and judgment selection, not a property of the query itself. TREC recall figures are best treated as range indicators, not absolute measurements.

## 7.5 Gaps in Digital Forensics

No standardized benchmark for keyword search or TAR performance at corpus scale exists in the digital forensics literature. Available evidence is limited to:

- A single study of ML-based ranking of forensic string search hits (Beebe & Liu, 2014).
- Emerging work on ML triage of specific artifact types (web browsing, session logs).

No equivalent of the TREC Legal Track has been conducted on forensic disk image corpora under reproducible conditions. Performance figures from eDiscovery benchmarks must serve as the best available proxy for general document classification performance in forensic contexts, with appropriate caution about differences in corpus type, vocabulary, and relevance criteria.

## 7.6 Generalizability of Best-Case Results

The exceptional keyword performance reported for H5's engineered deterministic query (76.2% recall, 84.4% precision, F1 80.1%) and for the TREC Total Recall Track hybrid workflow (F1 82%–100%) represent best-case outcomes under benchmark conditions with dedicated measurement infrastructure and expert practitioners. These figures are not generalizable to typical production environments and should not be used as representative baselines.

# 8. Final Conclusions

## 8.1 Defensible Benchmark Ranges

The following ranges represent defensible baselines derived from multiple consistent, well-evidenced sources. They are appropriate for use as reference points in comparative evaluations of classification systems.

### Keyword Search

| Metric | Defensible Range | Confidence | Key Condition |
|---|---|---|---|
| Recall (typical production) | **20%–40%** | High | Iterative Boolean/keyword search in legal review |
| Recall (reference Boolean query, TREC) | **4%–24%** | Moderate | Variable due to judgment methodology |
| Precision (narrow queries, low recall) | **70%–79%** | High | Query restricted to maximize precision at cost of recall |
| Precision (broad queries, moderate recall) | **10%–30%** | High | Query broadened to increase recall |
| F1 (typical operating points) | **~25%–40%** | High | Computed from above recall/precision pairs |
| Miss rate | **60%–80%** | High | At 20%–40% recall |

**Condition for validity:** These ranges apply to Boolean/keyword search without secondary machine learning classification, on collections of at least tens of thousands of documents with mixed vocabulary (e.g., email corpora, mixed-custodian litigation sets). Highly structured collections with predictable terminology may yield higher recall; informal communications typically yield lower recall.

### TAR — TAR 1.0 (SPL/SAL)

| Metric | Defensible Range | Confidence | Key Condition |
|---|---|---|---|
| Recall | **50%–75%** | Moderate | Standard passive or simple active learning protocol |
| Precision | **60%–80%** | Moderate | Well-calibrated model with quality training data |
| F1 | **~55%–75%** | Moderate | Computed from above |

### TAR — TAR 2.0 (CAL)

| Metric | Defensible Range | Confidence | Key Condition |
|---|---|---|---|
| Recall | **75%–96%** | High | CAL with consistent human review throughout |
| Precision | **80%–96%** | High | Same conditions |
| F1 | **~75%–96%** | High | Computed from above |
| Review effort reduction | **50%–80%** | Moderate | Relative to exhaustive manual review or broad keyword culling |

**Condition for validity:** CAL ranges are well-established for large email collections (Enron FERC, Jeb Bush, Kaine gubernatorial corpora) with prevalence in the range of 1%–10%. Performance may be lower in very low-prevalence environments (<0.5% relevant) or with inconsistent SME training labels.

### Human Linear Review (Reference Baseline)

| Metric | Defensible Range | Confidence |
|---|---|---|
| Recall | **49%–54%** | Moderate |
| Precision | **18%–20%** | Moderate |
| F1 | **~27%–28%** | Moderate |

Human linear review is not a reliable benchmark and consistently performs below TAR 2.0 at both recall and precision.

## 8.2 Confidence Qualifications

1. **Blair & Maron keyword recall (20%):** Highest confidence single data point in the literature. Foundational field study replicated conceptually across multiple subsequent evaluations.
2. **CAL recall/precision (75%–96%):** High confidence. Supported by multiple peer-reviewed studies (SIGIR 2014, SIGIR 2017), TREC Legal Track data, and the TREC Total Recall Track, across multiple large email collections.
3. **TREC reference Boolean query recall (4%–24%):** Moderate confidence. These figures are sampling-based estimates sensitive to judgment methodology. They should be treated as illustrative of the keyword performance range rather than fixed point estimates.
4. **TAR 1.0 (SPL/SAL) recall (50%–75%):** Moderate confidence. Fewer head-to-head controlled comparisons than CAL. Subject to wide variation depending on seed set quality and stopping criteria.
5. **Human review baseline (49%–54% recall, 18%–20% precision):** Moderate confidence. Based primarily on Roitblat et al. (2010), which measured against an imperfect original review. These figures represent the only controlled measurement of unaided human linear review on a large email corpus; actual performance against an ideal gold standard may differ.

6. **Forensic ML triage benchmarks:** Limited confidence for generalization. Single study (Beebe & Liu, 2014) on a synthetic disk image. No replications on independent datasets.

### 8.3 The 65% Inter-Assessor Ceiling

Any comparative evaluation of classification systems against these baselines must account for the fact that human gold standards are inherently imperfect. Voorhees (2000) established that expert human assessors agree with each other on relevance approximately 65% of the time. A classification system that achieves measured accuracy above ~65%–70% relative to a single human assessor's judgments may be correctly identifying documents that the assessor missed—but will be penalized as false positives within the evaluation framework. This ceiling applies equally to keyword search, TAR, and any newer classification method, and should be explicitly acknowledged when interpreting benchmark comparisons.

---

## 9. Reference Sources


1. **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. https://dl.acm.org/doi/10.1145/3166.3197
2. **Voorhees, E.M.** (2000). "Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness." *Information Processing & Management*, 36(5), 697–716.
3. **Baron, J.R.** (2005–2007). "Toward a Federal Benchmarking Standard for Evaluating the Retrieval of Electronic Information in Discovery." *The Sedona Conference Journal*, 6. https://www.thesedonaconference.org/sites/default/files/publications/237-246%20Baron_237-246%20Baron.qxd__0.pdf
4. **Tomlinson, S., Oard, D.W., Baron, J.R. & Thompson, P.** (2008). "Overview of the TREC 2007 Legal Track." *Proceedings of TREC 2007*, NIST Special Publication SP 500-274. https://trec.nist.gov/pubs/trec16/papers/LEGAL.OVERVIEW16.pdf
5. **Oard, D.W., Hedin, B., Tomlinson, S. & Baron, J.R.** (2009). "Overview of the TREC 2008 Legal Track." *Proceedings of TREC 2008*. https://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf
6. **Hedin, B., Tomlinson, S., Baron, J.R. & Oard, D.W.** (2010). "Overview of the TREC 2009 Legal Track." *Proceedings of TREC 2009*. https://trec.nist.gov/pubs/trec18/papers/LEGAL09.OVERVIEW.pdf
7. **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://onlinelibrary.wiley.com/doi/abs/10.1002/asi.21233
8. **Grossman, M.R. & Cormack, G.V.** (2011). "Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review." *Richmond Journal of Law & Technology*, XVII(3), Art. 11. https://scholarship.richmond.edu/cgi/viewcontent.cgi?article=1344&context=jolt
9. **Cormack, G.V. & Grossman, M.R.** (2014). "Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery." *SIGIR '14: Proceedings of the 37th International ACM SIGIR Conference*. https://dl.acm.org/doi/10.1145/2600428.2609601

10. **Cormack, G.V. & Grossman, M.R.** (2015). “Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review.” arXiv preprint. https://arxiv.org/pdf/1504.06868
11. **Grossman, M.R. & Cormack, G.V.** (2013). “The Grossman-Cormack Glossary of Technology-Assisted Review.” *Federal Courts Law Review*, 7(1), 1–34.
12. **Grossman, M.R. & Cormack, G.V.** (2017). “Navigating Imprecision in Relevance Assessments on the Road to Total Recall: Roger and Me.” *SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference*. https://dl.acm.org/doi/10.1145/3077136.3080812
13. **Losey, R.** (2015, 2016). “e-Discovery Team at TREC 2015 Total Recall Track.” TREC report. https://trec.nist.gov/pubs/trec24/papers/eDiscoveryTeam-TR.pdf
14. **Beebe, N.L. & Liu, L.** (2014). “Ranking Algorithms for Digital Forensic String Search Hit Triage.” *Digital Investigation*, 11(S1), S78–S87. (Presented at DFRWS 2014.) https://dfrws.org/sites/default/files/session-files/2014_USA_paper-ranking_algorithms_for_digital_forensic_string_search_hits.pdf
15. **Tomlinson, S.** (2007–2009). Series: “Experiments with the Negotiated Boolean Queries of the TREC Legal Track.” *TREC Proceedings*. https://apps.dtic.mil/sti/pdfs/ADA517742.pdf
16. **Oard, D.W. & Webber, W.** (2013). “Information Retrieval for E-discovery.” *Foundations and Trends in Information Retrieval*, 7(2–3), 99–237.
17. **Dimm, B.** (2018). “Keyword vs. CAL Results.” Clustify blog. https://blog.cluster-text.com/2018/12/ *(Industry study; not peer-reviewed.)*

# Appendix: Master Comparison Table

| Dimension | Keyword Search | TAR 1.0 (SPL/SAL) | TAR 2.0 (CAL) | Human Linear Review |
|---|---|---|---|---|
| **Typical recall** | 20%–40% | 50%–75% | 75%–96% | 49%–54% |
| **Typical precision** | 10%–79% (query-dep.) | 60%–80% | 80%–96%† | 18%–20% |
| **Typical F1** | ~25%–40% | ~55%–75% | ~75%–96% | ~27%–28% |
| **Miss rate** | 60%–80% | 25%–50% | 4%–25% | 46%–51% |
| **Evidence quality** | Well-established | Moderately supported | Well-established | Moderately supported |
| **Primary corpus type in evidence** | Legal / email / mixed | Email / legal | Email / legal | Email / legal |
| **Primary sources** | Blair & Maron 1985; TREC Legal Track 2007–2009 | TREC 2009; Grossman & Cormack 2011 | Cormack & Grossman 2014, 2017; TREC Total Recall 2015–2016 | Roitblat et al. 2010; Voorhees 2000 |

**Universal caveat:** All ranges reflect documented operating points from controlled or semi-controlled studies. Real-world results vary based on dataset size, prevalence, query or training quality, and stopping criteria. The ~65% inter-assessor agreement limit (Voorhees, 2000) constrains the interpretability of any method's measured performance against a single human gold standard.

†Precision ranges for TAR 2.0 reflect benchmark datasets with moderate-to-high prevalence (typically >1%–2%). In low-prevalence environments (<1% relevant), observed precision may be substantially lower even with high recall, due to base-rate effects (see §4.5 and §7.3).

---

*All figures in this document are sourced to citable primary or secondary literature or computed directly from reported metrics (e.g., F1 derived from reported recall and precision; miss rate derived as 1 − recall). Aggregated ranges are synthesized across multiple studies. Computed or approximate values are indicated by the ~ prefix. No values have been invented. Where a finding is supported by only one source, this is stated explicitly, and the finding is not used to establish a typical range.*