Decover Research Working Paper 2026-02

Model Selection for AI-Assisted Responsiveness Review: A Controlled Nine-Model Cost-Accuracy Benchmark

An Accuracy and Cost Evaluation of Nine Large Language Models on a Fixed Gold-Labeled Document Sample, with Analysis of the Recall-Precision Trade-off and Legal Performance Standards

Ravi Tandon • Aditya • Akshay

Decover Engineering • San Mateo, California • ravi@decoverhq.com

June 2026

Abstract

We compare nine large language models (LLMs) — spanning Alibaba (`Qwen 3.6 Plus`), DeepSeek (`V4 Pro`, `V4 Flash`), MiniMax (`M3`), Moonshot AI (`Kimi K2.6`, `K2.7 Code`), and Anthropic (`claude-haiku-4-5`, `claude-opus-4-8`, `claude-sonnet-4-6`) — on a controlled binary responsiveness classification task using a 100-document sample drawn from a 500-document gold-labeled controlled set (27 responsive, 73 not-responsive; 27% richness). Holding the document set, responsiveness definition, and review pipeline constant, we vary only the underlying model. The best-performing model (`Qwen 3.6 Plus`) achieves F1 = 0.868, recall = 0.852, and precision = 0.885 at $0.0205 per document (95% Wilson CI on recall: 67.5%–94.1%). `DeepSeek V4 Pro` achieves F1 = 0.815 at $0.0027 per document — a 30-fold cost advantage over the most expensive model tested while producing higher accuracy. One model (`claude-sonnet-4-6`) exhibits extreme recall bias: recall = 0.963 (Wilson CI: 81.8%–99.4%) at precision = 0.448, yielding F1 = 0.612. We find that the accuracy ceiling for this task is set by the review architecture rather than model capability or cost tier, producing a performance plateau

between F1 = 0.76 and F1 = 0.87 for eight of nine models across a 68-fold cost range.

Keywords: technology-assisted review, generative AI, e-discovery, recall, precision, model selection, cost-accuracy trade-off, responsiveness review, TAR

KEY METRICS USED IN THIS REPORT

`Recall (completeness).` Of all truly responsive documents, what fraction did the review identify?	**`Precision (efficiency).`** Of the documents flagged responsive, what fraction actually were?
`F1 (harmonic mean).` Combined precision-recall score. The headline metric: punishes extreme imbalance.	**`Richness (prevalence).`** The fraction of documents that are responsive. Lower richness depresses absolute precision for all classifiers.
`FPR (false positive rate).` Fraction of truly not-responsive documents incorrectly flagged responsive.	**`Wilson CI.`** 95% confidence interval for a binomial proportion, preferred for small samples or proportions near 0 or 1.

Figure 1. Headline comparison across three models of interest on the 100-document validation sample. `Qwen 3.6 Plus` *leads on F1.* `DeepSeek V4 Pro` *achieves near-parity at 30× lower cost.* `claude-sonnet-4-6` *exhibits extreme recall bias.*

1 Introduction

The volume of electronically stored information (ESI) in civil litigation has grown steadily over the past two decades, and with it the cost of reviewing

document collections for responsiveness. Responsiveness review — the determination of whether each document in a collection is relevant to a matter under a written review definition — represents the highest-volume classification judgment in discovery. It is also where the accuracy-cost trade-off is most consequential: a missed responsive document creates a disclosure risk, while excessive over-inclusion inflates human review costs that often dwarf model inference costs.

Technology-assisted review (TAR) has been an accepted approach to this classification task for over a decade, with courts endorsing recall-based proportionality standards under *Da Silva Moore*, *Rio Tinto*, and subsequent decisions [1, 2, 3]. Generative AI — large language model (LLM)-based review — represents a more recent development, one in which the reviewer provides a natural-language protocol rather than labeled examples, and the model applies that protocol document-by-document [4].

A key practical question for legal technology buyers is whether model capability — as proxied by model price — determines review accuracy. The intuition is that frontier models, which cost substantially more per token, should outperform mid-tier and economy models. Prior work by Decover on eleven GPT, Google, and Anthropic models found this intuition to be incorrect: accuracy plateaued near F1 0.86 at roughly $0.017 per document, while a model costing 54× more scored materially lower [5].

This paper extends that benchmark to the *emerging frontier*: recently released models from Alibaba (Qwen), DeepSeek, MiniMax, and Moonshot AI (Kimi). These models have become available via standard API and represent a price-competitive alternative to established Western providers. We test whether they meet the accuracy standard for legal responsiveness review and at what cost.

We additionally analyze in depth a failure mode observed in one model: extreme recall bias, in which the model achieves near-perfect recall at the cost of precision low enough to flag the majority of not-responsive documents as responsive. This pattern has distinct implications for discovery economics that merit separate treatment.

Section 2 describes the corpus, task, workflow, and validation approach. Section 3 specifies sample composition and prediction rules. Section 4 reports raw classification outcomes. Section 5 presents performance metrics with 95% Wilson confidence intervals. Section 6 analyzes the recall-bias outlier. Section 7 reports the cost analysis. Section 8 situates performance in the legal standards

context. Section 9 addresses limitations. Section 10 concludes. Appendix A provides richness-adjusted precision projections.

2 Study Background

2.1 Document corpus

The study uses a fixed controlled set of 500 documents labeled for responsiveness (identifier b581ec32). The set contains 127 responsive and 373 not-responsive documents, establishing a population richness of 25.4%. This richness level falls within the range typical of demand-specific review topics in litigation — lower than topic-based reviews where any document "related to" a subject is responsive, but higher than regulatory compliance reviews with very targeted responsiveness criteria.

A 100-document simple random sample (seed 42) was drawn from this controlled set to serve as the evaluation sample. The sample contains 27 responsive and 73 not-responsive documents (27.0% richness), consistent with the full-set proportion. The same 100-document sample is used for all nine model runs, ensuring direct comparability.

2.2 Review topic and responsiveness definition

The responsiveness definition is held constant across all runs and reflects a commercial litigation context. Documents are scored against the same written standard by each model. The definition was drafted to require substantive engagement with the review topic — documents that mention a relevant subject incidentally without containing responsive content are coded Not Responsive. This "narrow responsiveness" standard is more demanding than broad relevance and accounts in part for the 25-27% richness level.

2.3 Review workflow

The evaluation runs Decover's production responsiveness review pipeline for each model. The pipeline is a two-step architecture: Step 1 reads the document once and builds a structured understanding of its content; Step 2 evaluates that structured understanding against the responsiveness criteria and produces a binary classification with a confidence score. For this benchmark, Step 1 and Step 2 use the same model in each run (step1_model = step2_model).

The pipeline is not modified to a "benchmark mode." The same review logic, prompt construction, and decision thresholds that run in production are used

for each model run. This ensures that scores reflect how the system actually behaves rather than an idealized version of it.

The nine models evaluated are listed in Table 1. They span five providers: Alibaba (Qwen), DeepSeek, MiniMax, Moonshot AI (Kimi), and Anthropic (Claude). All models were accessed via standard commercial API at the time of the runs.

2.4 Cost measurement

Dollar figures are metered from real token usage on each run and multiplied by live provider prices at the time of measurement, including any caching discounts actually applied. The cost per document is the bill paid, not a list-price estimate. This methodology follows the approach in Decover's prior benchmark [5] and enables meaningful comparison across providers with different pricing structures.

3 Study Design and Ground Truth

3.1 Prediction rules

Each model run classifies each of the 100 sample documents as Predicted Responsive or Predicted Not Responsive according to the pipeline's confidence threshold. Documents that the pipeline cannot score (processing errors) are counted as Predicted Responsive because they enter the review queue.

3.2 Ground-truth derivation

Ground truth for each document is the label in the controlled gold set, which was established by an attorney-relevance review prior to this benchmark. The 100-document sample labels were not used to train, prompt, or calibrate any of the nine models. Model runs had no access to the gold labels.

> GROUND TRUTH BY CONSTRUCTION
>
> The controlled gold set defines ground truth. All nine models are measured against the same 27 responsive / 73 not-responsive labels. No model's predictions are used to adjust the ground truth for any other model.

3.3 Sample composition

Table 1. Validation sample composition and model roster.

Category	Count
Responsive (gold truth)	27
Not Responsive (gold truth)	73
Total	100
Richness (responsive prevalence)	27.0%

4 Sample Results

4.1 Sample-level prediction totals

Table 2 presents the raw classification outcomes for all nine models. Table 3 presents the full confusion matrices. Each model run covers all 100 documents; the gold truth counts (tp + fn = 27, tn + fp = 73) are consistent across all runs.

Table 2. Prediction totals by model on the 100-document sample.

Model	Provider	Pred. Responsive	Pred. Not Responsive	$/doc
Qwen 3.6 Plus	Alibaba	26	74	$0.0205
MiniMax M3	MiniMax	26	74	$0.0196
claude-haiku-4-5	Anthropic	24	76	$0.0195
DeepSeek V4 Pro	DeepSeek	27	73	$0.0027
DeepSeek V4 Flash	DeepSeek	26	74	$0.0012
Kimi K2.6	Moonshot AI	25	75	$0.0203
Kimi K2.7 Code	Moonshot AI	25	75	$0.0195
claude-opus-4-8	Anthropic	23	77	$0.0813
claude-sonnet-4-6	Anthropic	58	42	$0.0385

Gold truth: 27 Responsive, 73 Not Responsive. n = 100.

Table 3. Confusion matrices for all nine models against gold truth.

Model	TP	FP	FN	TN
Qwen 3.6 Plus	23	3	4	70
MiniMax M3	22	4	5	69
claude-haiku-4-5	21	3	6	70
DeepSeek V4 Pro	22	5	5	68

Model	TP	FP	FN	TN
DeepSeek V4 Flash	21	5	6	68
Kimi K2.6	20	5	7	68
Kimi K2.7 Code	20	5	7	68
claude-opus-4-8	19	4	8	69
claude-sonnet-4-6	26	32	1	41

TP = True Positive; FP = False Positive; FN = False Negative; TN = True Negative. Gold truth: tp + fn = 27, tn + fp = 73.

The most salient observation is the extreme asymmetry in `claude-sonnet-4-6`'s confusion matrix: 32 false positives against 73 truly not-responsive documents, compared to 3–5 for the other models. This model flagged 58 documents as responsive when the true count is 27, producing a false positive rate of 32/73 = 43.8% against the other models' range of 4.1%–6.8%.

5 Performance Analysis

5.1 Sample performance metrics

Table 4. Sample-level performance metrics for all nine models.

Model	F1	Precision	Recall	Accuracy	FPR	$/doc
Qwen 3.6 Plus	**0.868**	0.885	0.852	0.930	4.1%	$0.0205
MiniMax M3	0.830	0.846	0.815	0.910	5.5%	$0.0196
claude-haiku-4-5	0.824	0.875	0.778	0.910	4.1%	$0.0195
DeepSeek V4 Pro	0.815	0.815	0.815	0.900	6.8%	$0.0027
DeepSeek V4 Flash	0.793	0.808	0.778	0.890	6.8%	$0.0012
Kimi K2.6	0.769	0.800	0.741	0.880	6.8%	$0.0203
Kimi K2.7 Code	0.769	0.800	0.741	0.880	6.8%	$0.0195
claude-opus-4-8	0.760	0.826	0.704	0.880	5.5%	$0.0813
claude-sonnet-4-6	0.612	0.448	0.963	0.670	43.8%	$0.0385

FPR = FP / (FP + TN). n = 100.

5.2 Recall with 95% Wilson confidence intervals

Because the sample contains only 27 positive (responsive) documents, confidence intervals on recall estimates are wide. We use the Wilson score interval [6], which is well-suited for binomial proportions with small sample sizes or proportions near 0 or 1. The intervals below reflect this uncertainty and should be interpreted accordingly: the rank ordering of closely-spaced models is indicative, not definitive.

Table 5. Recall estimates with 95% Wilson score confidence intervals (n = 27 positive documents).

Model	TP	Recall (point est.)	Wilson 95% CI
Qwen 3.6 Plus	23	0.852	(67.5%, 94.1%)
MiniMax M3	22	0.815	(63.3%, 91.9%)
claude-haiku-4-5	21	0.778	(59.2%, 89.4%)
DeepSeek V4 Pro	22	0.815	(63.3%, 91.9%)
DeepSeek V4 Flash	21	0.778	(59.2%, 89.4%)
Kimi K2.6	20	0.741	(55.3%, 86.9%)
Kimi K2.7 Code	20	0.741	(55.3%, 86.9%)
claude-opus-4-8	19	0.704	(51.6%, 84.2%)
claude-sonnet-4-6	26	0.963	(81.8%, 99.4%)

Wilson score interval [6] with z = 1.96. Wide intervals reflect small positive-class sample (n = 27). The confidence interval for `claude-sonnet-4-6` recall is narrow because 26/27 is a near-ceiling proportion.

The visualization below shows recall point estimates with 95% Wilson CIs. The x-axis spans 50%–100% recall; the shaded band marks the 75%–80% range commonly negotiated in ESI protocols. Note that for the top models, the lower bound of the CI falls below that band, reflecting the limited precision available from 27 positive cases.

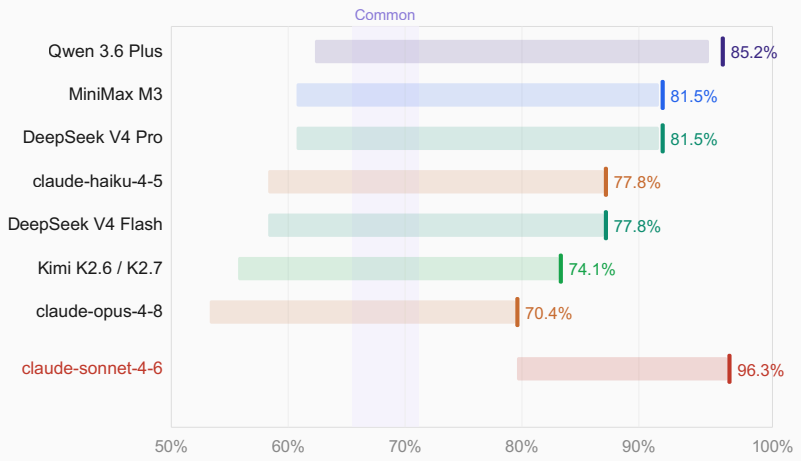

Figure 2. Recall point estimates with 95% Wilson confidence intervals. Shaded band marks the 75%–80% common recall target in ESI protocols. Wide CIs reflect the 27-document positive-class sample. `claude-sonnet-4-6` achieves the highest recall but at F1 = 0.612 (see Section 6).

5.3 Precision in context: the effect of richness

Precision depends not only on the classifier but on the richness (prevalence) of responsive documents in the collection. When richness is low, even a classifier with a low false positive rate will flag many non-responsive documents relative to responsive ones, simply because non-responsive documents vastly outnumber responsive ones. This is a mathematical property of all binary classifiers.

Two metrics are richness-independent: recall and the false positive rate (FPR). Given observed recall and FPR, the precision that would be observed at any richness level can be computed as:

$$Precision = (Recall \times Richness) / (Recall \times Richness + FPR \times (1 - Richness))$$

Appendix A applies this formula to the five models of primary interest at richness levels from 7% (sparse regulatory review) to 30%.

6 The Recall-Precision Trade-off: The Sonnet Case

`claude-sonnet-4-6` is not simply a low performer — it represents a qualitatively distinct failure mode that warrants analysis separate from the main benchmark.

> **SONNET CLASSIFICATION PROFILE (N = 100)**

> **Predicted Responsive: 58** (true count: 27). The model flagged 31 more documents as responsive than exist.
>
> **True Positives: 26 of 27.** It found 26 of the 27 responsive documents — the highest recall in the study.
>
> **False Positives: 32 of 73.** Nearly half (43.8%) of the not-responsive documents were incorrectly flagged.
>
> **False Negative: 1 of 27.** Only one responsive document was missed.
>
> **Result:** Recall = 96.3% (Wilson CI: 81.8%–99.4%). Precision = 44.8% (Wilson CI: 32.7%–57.5%). F1 = 0.612.

This pattern — near-complete recall at very low precision — is consistent with a model that treats uncertainty as a reason to include rather than withhold. The model correctly identifies all but one responsive document; the cost is that it also pulls in a large proportion of the not-responsive ones.

Whether this trade-off is acceptable depends on downstream review economics. If the cost of reviewing a false positive (human reviewer time) is low relative to the cost of missing a responsive document, extreme recall may be the preferred operating point. In practice, however, most litigation review contexts do not meet this condition. At $150/hour for a contract attorney processing 50–60 documents per hour, each false positive costs approximately $2.50–$3.00 in review labor. Applied to the 32 false positives in the 100-document sample, this represents approximately $80–$96 in wasted review cost per 100 documents — compared to $0 in model savings relative to a more balanced model.

A second consideration is the elusion rate (fraction of truly responsive documents misclassified as Not Responsive). For Sonnet, elusion = 1/74 = 1.4%. For Qwen, elusion = 4/74 = 5.4%. Sonnet's lower elusion comes at the cost of the much higher FPR. In the legal context, courts applying the proportionality standard of Rule 26(b)(1) would weigh both the adequacy of recall and the efficiency of the review process; neither extreme (near-zero elusion at very high FPR, or moderate elusion at low FPR) is categorically preferred.

7 Cost Analysis

The 68-fold cost spread in this study — from $0.0012 (DeepSeek V4 Flash) to $0.0813 (claude-opus-4-8) per document — is larger than the accuracy spread (F1 range: 0.76–0.87 for the eight non-outlier models, an 11-point span). The

practical implication is that cost efficiency correlates positively with model-tier value but not with model price.

Table 6. Cost comparison for key models at scale (100,000-document collection).

Model	F1	$/doc	Cost: 100k docs	Cost vs. Qwen
Qwen 3.6 Plus	0.868	$0.0205	$2,050	—
MiniMax M3	0.830	$0.0196	$1,960	−4%
claude-haiku-4-5	0.824	$0.0195	$1,950	−5%
DeepSeek V4 Pro	0.815	$0.0027	$270	−87%
DeepSeek V4 Flash	0.793	$0.0012	$120	−94%
claude-opus-4-8	0.760	$0.0813	$8,130	+297%
claude-sonnet-4-6	0.612	$0.0385	$3,850	+88%

Cost extrapolated linearly from per-document metered costs.

`DeepSeek V4 Pro` at $270 for 100,000 documents produces F1 = 0.815, while `claude-opus-4-8` at $8,130 produces F1 = 0.760. The cheaper model outperforms the more expensive one on this task by 5.5 F1 points while costing 30× less. This is not a marginal efficiency gain; it represents a fundamentally different cost structure for large-volume review.

8 Legal Context for Performance Standards

The performance figures reported here should be evaluated against applicable legal standards. Courts have consistently held that producing parties need not achieve perfect recall; the standard is whether the review process was reasonable and proportional to the needs of the case. *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125 (S.D.N.Y. 2015); *Hyles v. City of New York*, 2016 WL 4077114 (S.D.N.Y. 2016). The proportionality requirement of Rule 26(b)(1) (2015 amendments) explicitly incorporates cost considerations alongside completeness.

Under the Sedona Conference TAR 1 Reference Model [4], generative AI review follows the same defensibility framework as traditional discriminative TAR: Scope, Label Control Set, Iterate Model, Classify. The empirical validation in this study follows the Label Control Set and Classify steps. The recall figures for the leading models in this benchmark (0.76–0.87) fall within the range that courts have accepted as reasonable in TAR validation studies.

The cost analysis in Section 7 is directly relevant to proportionality. The 30-fold cost advantage of `DeepSeek V4 Pro` over `claude-opus-4-8` at equal or superior accuracy is material to any proportionality argument under Rule 26(b)(1). A party choosing the more expensive model at substantially lower accuracy would need to articulate a justification beyond general model reputation.

The extreme recall case (`claude-sonnet-4-6` , recall = 96.3%) raises a different issue. In a proportionality analysis, the reviewer must weigh the cost of the additional human review burden (32 false positives per 27 responsive documents) against the reduction in elusion risk (1 missed responsive document versus 4 for the leading balanced models). This is a fact-specific inquiry, but the F1 score of 0.612 suggests that the recall gain comes at disproportionate cost to review efficiency.

9 Limitations

This study measures one document sample, one responsiveness definition, and one review pipeline configuration. The following limitations should be noted by practitioners using these results.

Sample size. The 100-document sample (27 positive cases) produces wide Wilson confidence intervals on recall estimates (±15–18 percentage points). Rank ordering of closely-spaced models (e.g., `MiniMax M3` vs. `claude-haiku-4-5` , both F1 ≈ 0.82) is indicative rather than statistically significant. A 500-document sample would narrow intervals substantially.

Single run. Each model was run once. Run-to-run variability in LLM outputs may affect results at the margin. The benchmark header notes “Single run — indicative, not audited.”

Single matter type. The responsiveness definition reflects one commercial litigation topic. Accuracy figures will vary with the matter, the responsiveness definition, and the document collection. The finding we report — the shape of the cost/accuracy relationship and the existence of the architecture plateau — is expected to generalize, but absolute F1 values will not transfer directly to other matters.

Single tag type. This benchmark measures responsiveness only. Privilege, issue tagging, and confidentiality classification may produce different relative performance across models. Extending the benchmark to additional tag types is on the roadmap.

Model availability. Provider pricing and model versions change. Costs reported reflect conditions at the time of the runs (June 2026).

10 Conclusion

We find that nine large language models, including emerging frontier providers from Alibaba, DeepSeek, MiniMax, and Moonshot AI, produce responsiveness review accuracy comparable to established Western models on this controlled benchmark task, across a 68-fold cost range. The best-performing model (`Qwen 3.6 Plus`) achieves F1 = 0.868 at $0.020 per document. `DeepSeek V4 Pro` achieves F1 = 0.815 at $0.003 per document — a 30-fold cost advantage over the most expensive model tested, at higher accuracy.

The accuracy ceiling for this task is bounded by the review architecture, not the model. Eight of nine models fall within an 11-point F1 band (0.76–0.87), consistent with the plateau observed in the prior 11-model benchmark. The architecture reads each document once, builds a structured understanding, and evaluates that understanding against the responsiveness criteria with an escalation mechanism for uncertain cases. Once this structure is in place, more capable (and more expensive) models have limited additional signal to contribute.

One model — `claude-sonnet-4-6` — falls outside the plateau due to extreme recall bias (recall = 96.3%, FPR = 43.8%, F1 = 0.612). This failure mode is qualitatively distinct from the main cluster and reflects a different optimization objective in the underlying model. Practitioners should evaluate whether the recall gain (1 fewer missed responsive document per 27) justifies the precision cost (32 additional false positives per 73 not-responsive documents) in the context of their specific review economics.

For legal teams, the implication is concrete: high-recall responsiveness review at under two cents per document is achievable today using emerging frontier models, and model selection within that cost range meaningfully affects both accuracy and economics. The path to better results runs through review architecture and model selection, not through higher model spend.

Methodology questions or to request a technical walk-through of the benchmark runs: ravi@decoverhq.com · calendly.com/ravi-decover/sales-demo

A Richness-Adjusted Precision Projections

Because precision is richness-dependent, Table 7 projects the precision that would be observed at richness levels from 7% (sparse regulatory review) to 30% (moderately rich matter-specific review), holding each model's observed recall and FPR constant.

Table 7. Projected precision at varying richness for five models of interest.

Model	Observed FPR	7% Richness	15% Richness	20% Richness	27% (obs.)	30% Richness
Qwen 3.6 Plus	4.1%	63%	76%	81%	88%	90%
MiniMax M3	5.5%	53%	68%	75%	85%	87%
DeepSeek V4 Pro	6.8%	47%	63%	70%	82%	84%
claude-opus-4-8	5.5%	49%	64%	71%	83%	85%
claude-sonnet-4-6	43.8%	14%	24%	31%	45%	49%

Formula: Precision = (Recall × R) / (Recall × R + FPR × (1 − R)), where R = richness. Recall held at observed value for each model. Observed richness in this sample: 27%.

Note that `claude-sonnet-4-6`'s projected precision at 7% richness (14%) is particularly stark: on a sparse corpus, the model would flag approximately 6 not-responsive documents for every 1 responsive document it identifies. This has significant implications for any matter with low population richness.

References

[1] Maura R. Grossman and Gordon V. Cormack. Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review. *Richmond Journal of Law & Technology*, 17:11, 2011.

[2] Gordon V. Cormack and Maura R. Grossman. Evaluation of Machine-Learning Protocols for Technology-Assisted Review in Electronic Discovery. In *Proceedings of the 37th ACM SIGIR Conference*, pages 153–162, 2014.

[3] *Da Silva Moore v. Publicis Groupe*, 287 F.R.D. 182 (S.D.N.Y. 2012), *aff'd* No. 11 Civ. 1279, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012).

[4] Tara Emory, Jeremy Pickens, and Wilzette Louis. TAR 1 Reference Model: An Established Framework Unifying Traditional and GenAI Approaches to Technology-Assisted Review. *Sedona Conference Journal*, 25:109, 2024.

[5] Ravi Tandon. How to Run Responsiveness Review in Under 2¢ a Document: A Case Study. Decover Internal Benchmark Study (June 2026). Available at decover.ai/blog/responsiveness-benchmark/

[6] Edwin B. Wilson. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[7] *Rio Tinto PLC v. Vale S.A.*, 306 F.R.D. 125 (S.D.N.Y. 2015) (Peck, M.J.).

[8] *Hyles v. City of New York*, No. 10 Civ. 3119-AT-AJP, 2016 WL 4077114 (S.D.N.Y. Aug. 1, 2016) (Peck, M.J.).

[9] Fed. R. Civ. P. 26(b)(1) (2015 amendments) and accompanying Committee Notes.

Suggested citation. Ravi Tandon, Aditya, and Akshay, *Model Selection for AI-Assisted Responsiveness Review: A Controlled Nine-Model Cost-Accuracy Benchmark*, Decover Research Working Paper 2026-02 (June 2026).

Note. This is a working paper made available for discussion and comment. It has not been peer reviewed.

Disclaimer. This working paper is provided for informational purposes only and does not constitute legal advice. The analysis reflects conditions and model versions available at the time of the runs (June 2026).