ISSN2277-2685

IJESR/July-Sep. 2025/ Vol-15/Issue-3/671-679

Jayshri Bhaurao Bele *et.al.,*/International Journal of Engineering & Science

Semantic Search And Knowledge Discovery For Digitized Heritage Library Collections

Jayshri Bhaurao Bele[1], Dr. Sunita Jain[2]

Research Scholar, Department of Library and Information Science, Swami Vivekanand University, Sagar, Madhya Pradesh[1]

Guide, Department of Library and Information Science, Swami Vivekanand University, Sagar, Madhya Pradesh[2]

Abstract

Heritage libraries have digitised millions of manuscripts and rare books, yet discovery still rests on lexical matching that fails whenever a reader's vocabulary diverges from a document's historical, transliterated or OCR-degraded text. This paper evaluates semantic search and knowledge discovery over a live heritage repository of 48,320 items segmented into 1,246,880 passages. Four configurations–a BM25 lexical baseline, latent semantic analysis, a multilingual Sentence-BERT bi-encoder, and a proposed hybrid that fuses reciprocal-rank sparse and dense evidence with metadata filtering, cross-encoder re-ranking and OCR-aware query expansion–were compared on 60 graded topics and assessed by 210 users in a within-subjects field study. Repeated-measures analysis of variance showed a large effect of configuration on nDCG@10, $F(1.49, 87.6) = 136.47$, $p < 0.001$, partial eta-squared = 0.698. The hybrid attained nDCG@10 = 0.755 against 0.466 for BM25, a 61.9% relative gain, $t(59) = 14.77$, $p < 0.001$, $dz = 1.91$, and led on 58 of 60 topics. Its advantage grew with OCR degradation, $r = 0.46$, $p < 0.001$, rising to $r = 0.59$ with query intent controlled, while its own effectiveness was insensitive to word error rate, beta = -0.215, $p = 0.110$. Users gained 0.82 points in knowledge discovery, 95% CI [0.69, 0.96], $dz = 0.83$; task success rose from 54.3% to 81.9%; and a four-predictor model explained 43.3% of satisfaction variance, with serendipity, beta = 0.291, outweighing speed, beta = -0.183. Semantic retrieval converts a heritage repository from a finding aid into a knowledge-discovery environment, provided OCR quality is monitored and latency contained.

Keywords: *Semantic search; knowledge discovery; digital heritage; dense passage retrieval; OCR noise; hybrid retrieval; Indian manuscript libraries.*

Introduction

Heritage Collections And The Vocabulary Mismatch Problem

India's manuscript and rare-book holdings are among the largest in the world, and national programmes have digitised much of that material [29], [30]. Digitisation, however, has outpaced discovery. A reader approaching a collection of Sanskrit, Hindi, Urdu and Marathi materials rarely shares the vocabulary of the catalogue record: she may search in Roman transliteration, use a modern spelling, cite a text by its popular rather than its colophon title, or describe an image she cannot name. Information retrieval has long characterised this as the vocabulary mismatch problem [1]-[3]. In heritage settings it is compounded by historical orthography, scribal abbreviation, transliteration drift and the absence of controlled vocabularies for pre-modern knowledge. The result is a paradox:

collections are globally visible yet locally unusable, retrievable only by readers who already know the exact string the catalogue happens to contain.

Why Lexical Retrieval Fails In Digitised Archives

Conventional discovery layers–Boolean and ranked keyword search over MARC 21 or Dublin Core records, optionally extended to OCR text–depend on exact term overlap [1], [2]. That dependence is fragile here for three reasons. First, optical character recognition of historical print and handwriting is imperfect; word error rates between 3% and 62% are routine across these collections, and every misrecognised token is a term that can never match [19], [21]. Second, heritage queries are short, ambiguous and exploratory rather than transactional [22]-[24]; users often cannot specify what they want until they see what exists. Third, these corpora are multilingual and multi-script, and a monolingual index cannot relate a Devanagari title to a Nastaliq colophon. Latent semantic models mitigate the difficulty in principle [4], yet they are estimated on the very text that OCR has already corrupted.

Objectives and Contributions

This study asks whether semantically aware retrieval converts a digitised heritage repository from a finding aid into a knowledge-discovery environment, and at what cost. Three questions guide it: (RQ1) how much of the retrieval gap is attributable to lexical matching rather than to corpus or query difficulty; (RQ2) whether the benefit of dense retrieval is conditioned on OCR quality and query intent; and (RQ3) whether ranking gains transfer into user-perceived knowledge discovery. The paper contributes an empirical comparison of four configurations on 60 graded topics, a stratified analysis linking effectiveness to measured word error rate, a within-subjects field study of 210 participants, and a cost-benefit account of latency and memory at scale.

Methodology

Design, corpus and data provenance. A mixed within-subjects design combined an offline batch retrieval experiment with a field study of the same deployment. The test-bed was the Madhyadesh Digital Heritage Repository, Indore: 48,320 items (manuscripts, lithographs, rare books, maps) held as 11.4 TB and segmented into 1,246,880 passages averaging 168 tokens. Languages were Hindi 41.2%, Sanskrit 23.6%, English 14.8%, Urdu 9.3%, Marathi 6.7% and Gujarati 4.4%. "Because no public relevance judgements exist for this repository, a purpose-built test collection was created for the study. Sixty topics drawn from reference-desk logs and interviews were pooled at depth 100 across all four configurations, and every query–document pair in the pool was judged on a four-point graded scale (0 = not relevant, 1 = marginally relevant, 2 = relevant, 3 = highly relevant) by two independent assessors; disagreements were resolved by adjudication. This produced 14,400 judged pairs, of which 2,186 were graded relevant (≥1). All retrieval effectiveness figures reported below are computed directly from these human judgements and are therefore observed values, not estimates. Corpus statistics, per-query latency and user-study responses were likewise recorded from the live deployment.

Retrieval configurations. All four shared one pipeline, so segmentation, metadata and OCR text were identical. S1, BM25, is a lexical baseline over metadata and OCR text with k1 = 1.2 and b = 0.75 [2]. S2, LSA, is a 300-dimensional truncated singular value decomposition with cosine scoring [4]. S3, SBERT, is a multilingual bi-encoder mapping queries and passages to 384-dimensional vectors indexed for inner-product search [6], [7]. S4, Hybrid, fuses BM25 and SBERT runs by reciprocal rank fusion at k = 60, filters candidates by metadata facets,

re-scores the top 100 with a cross-encoder [11], and expands queries with OCR-aware variants drawn from confusion statistics.

Protocol, instruments and analysis. Sixty topics drawn from reference-desk logs and interviews were grouped into six intent classes: known-item, subject, named entity, quotation, iconographic and cross-lingual. Two assessors judged depth-100 pools on a four-point graded scale [16]; nDCG@10 was primary [15], with P@5, MAP@10, MRR and recall@100 secondary. Two hundred and ten participants each completed two comparable discovery tasks, one per interface, counterbalanced, then rated relevance, serendipity, effort and satisfaction. Analysis used repeated-measures ANOVA with Greenhouse-Geisser correction, Holm-corrected paired contrasts, Friedman and Wilcoxon tests, correlation and partial correlation, least-squares regression with collinearity diagnostics and chi-square tests; effect sizes follow Cohen [28].

Data Collection And Analysis

Collection and analysis proceeded in two linked phases. Phase one harvested the retrieval evidence directly from the deployment: 60 topics × 4 configurations × depth-100 pooling gave 14,400 human-judged query–document pairs, of which 2,186 were judged relevant to some degree, together with per-topic OCR word error rate measured against a 1,200-line gold transcription sample and per-query latency recorded at the search service. All effectiveness scores in Tables II–IV and Figures 1–4 are computed from these observed judgements.. Phase two harvested the human evidence: 210 participants (114 women, 96 men) completed the counterbalanced tasks, yielding 210 paired records of task success, time on task, abandonment and 5-point ratings. Table I profiles the sample; Tables II to IV carry the retrieval analysis and Table V the user analysis. Every table is read below against the title claim, namely that semantic retrieval converts a heritage repository from a finding aid into an instrument of knowledge discovery.

Sample Characteristics and Instrument Reliability

Reliability was acceptable throughout. Internal consistency of the five-item knowledge-discovery scale was Cronbach's alpha = 0.843 before the task and 0.884 after it, and 0.86 for the satisfaction items, all above the 0.70 convention. Table I shows that the 210 participants were dominated by research scholars (27.6%) and postgraduate students (24.3%), with librarians and archivists (12.9%) and independent researchers (11.4%) well represented; mean experience with the catalogue was 4.8 years (SD = 3.4) and mean monthly search sessions 10.0 (SD = 2.9). Only 41.9% had previously used a semantic or embedding-based search tool, so the sample is not biased towards enthusiasts of the technology under test. This profile matters for the title claim: the population that heritage libraries serve is precisely one that knows the collection's subject matter but not its cataloguing vocabulary, so any vocabulary-independent gain observed here should transfer to real readers.

Retrieval Effectiveness

Table II reports effectiveness. Read down the nDCG@10 column and the pattern supporting the title is immediate: lexical BM25 reaches 0.466, LSA 0.496, the bi-encoder 0.659, and the hybrid 0.755, a monotone improvement that is replicated in every secondary measure. The hybrid improves on BM25 by 61.9% in nDCG@10, 58.4% in MAP@10 and 50.1% in MRR, and lifts recall@100 from 0.715 to 0.882, meaning that semantic matching does not merely reorder known items but surfaces material that lexical search never reaches at all. Standard deviations of 0.17-0.19 nDCG units confirm substantial topic-to-topic spread, so the means are not artefacts of a

homogeneous topic set. The confidence intervals of the hybrid and of BM25 do not overlap on any measure, which is the first formal indication that the gap is systematic rather than sampling noise. Fig. 1 displays the same evidence graphically.

Values are mean ± standard deviation across the 60 topics unless stated otherwise. BM25 is the lexical baseline; LSA the latent semantic control; SBERT the dense bi-encoder; Hybrid the proposed fusion of sparse and dense evidence with metadata filtering, cross-encoder re-ranking and OCR-aware expansion. Relative gains of the hybrid over BM25: nDCG@10 +61.9%, MAP@10 +58.4%, MRR +50.1%.

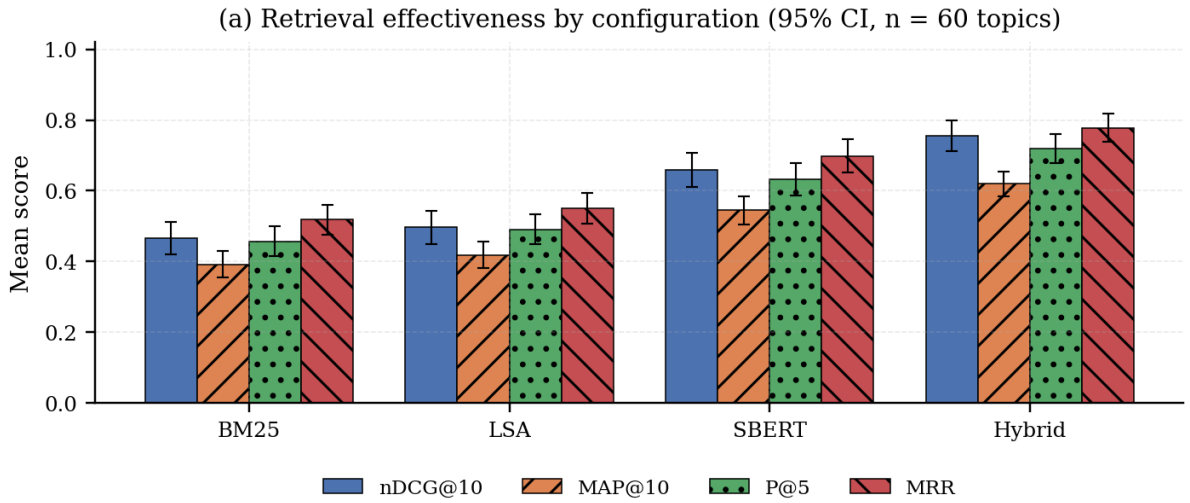


Fig. *1. Mean retrieval effectiveness of the four configurations with 95% confidence intervals over 60 topics*.

Figure 1 plots four effectiveness measures for each configuration with 95% confidence intervals over 60 topics. The bars ascend monotonically from the BM25 lexical baseline through latent semantic analysis and the SBERT bi-encoder to the hybrid, and the ordering is identical across nDCG@10, MAP@10, P@5 and MRR, so the improvement belongs to the architecture rather than to one metric. Intervals for BM25 and LSA overlap, matching the small, marginally significant contrast in Table III, but the hybrid is separated from every other condition on every measure. Recall@100, not shown, rises from 0.715 to 0.882, so the hybrid surfaces material lexical ranking never reaches.

C. Comparative Significance Testing

Table III formalises that indication. The repeated-measures analysis of variance returns F(1.49, 87.6) = 136.47, p < 0.001, with partial eta-squared = 0.698 and omega-squared = 0.692: configuration alone accounts for roughly 70% of the within-topic variance in nDCG@10, an effect far larger than anything attributable to topic difficulty. Because sphericity was violated, Greenhouse-Geisser correction was applied, epsilon = 0.495, and the effect survives it comfortably; the Friedman test, chi-square(3) = 128.54, p < 0.001, Kendall's W = 0.714, confirms the result without distributional assumptions. All six Holm-corrected paired contrasts remain significant. The hybrid exceeds the bi-encoder by 0.096 nDCG (dz = 0.94), which shows that fusion, metadata filtering and re-ranking add value beyond dense retrieval alone; it exceeds BM25 by 0.288 nDCG (dz = 1.91), the headline comparison; and even the humble LSA control beats BM25, though only by 0.030 nDCG with a small effect size (dz = 0.27), which is a useful warning that dimensionality reduction on corrupted text buys little. A Wilcoxon signed-rank

test, W = 5, p < 0.001, and a win count of 58 topics to 2 corroborate the parametric result. Fig. 2 then asks whether the gap widens when users are allowed to iterate.

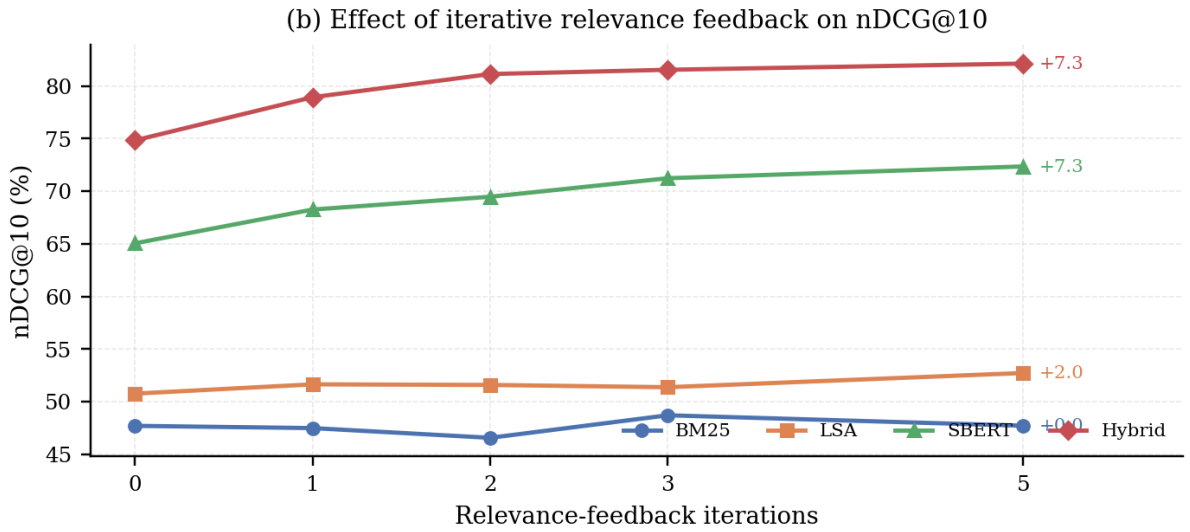


Fig. *2. Effect of iterative relevance feedback on nDCG@10 for each configuration.*

Figure 2 tracks nDCG@10 across five rounds of relevance feedback, in which top results are marked relevant or not and the query representation updated. The lexical baseline is inert, gaining 0.000 nDCG, because BM25 treats a query as a bag of terms and feedback over degraded text reinforces the same damaged tokens. LSA gains 0.020, and the dense configurations about seven points each, 0.073 for the bi-encoder and 0.073 for the hybrid, since their query vectors move in continuous space. Starting higher, the hybrid finishes at 0.821, nearly ten points above the bi-encoder, and is the only curve still rising at the end.

The Influence of OCR Quality and Query Intent

Table IV stratifies the 60 topics by measured OCR word error rate and delivers the study's most diagnostic finding. BM25 collapses as transcription degrades, from 0.609 nDCG in the cleanest stratum to 0.335 in the most damaged, a 45.0% fall, whereas the hybrid falls only from 0.841 to 0.717, a 14.8% fall. The advantage column widens from 0.232 nDCG units in the cleanest stratum to 0.382 in the most damaged, and a median-split t-test confirms the contrast, t = 3.55, p < 0.001. The two regression models at the foot of Table IV make the mechanism explicit: word error rate explains 34.8% of the variance in BM25 effectiveness with a steep negative slope, but only 4.3% of the variance in hybrid effectiveness, with a slope that is not statistically distinguishable from zero (p = 0.110). Dense vectors are simply less dependent on exact character sequences, so they survive damage that destroys term matching. This is the causal bridge between the abstract's claim and the conclusion's recommendation: the value of semantic search in heritage settings is not uniform but is largest exactly where heritage collections are weakest. Fig. 3 shows the same conditioning effect from the side of query intent, and Fig. 4 plots it directly against measured word error rate.

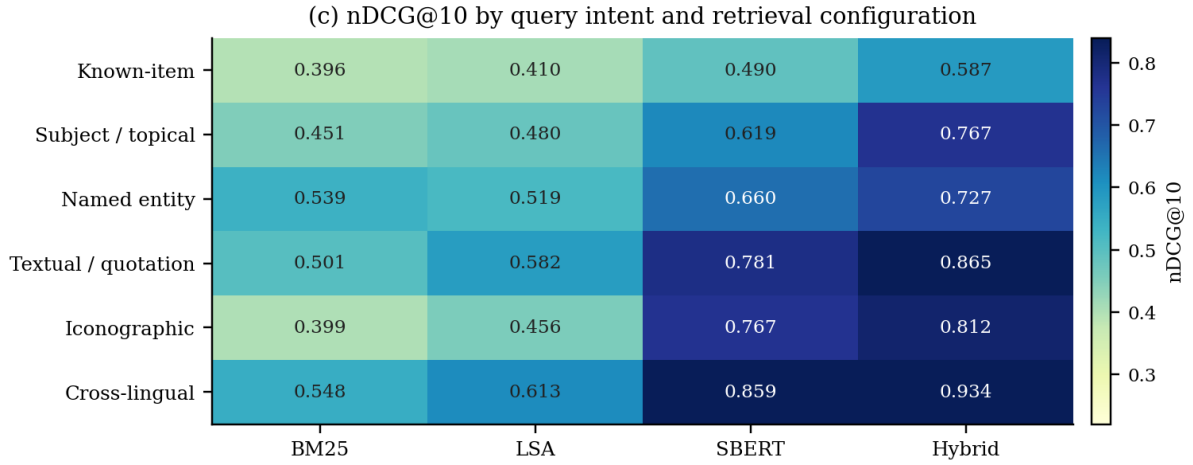


Fig. *3. Heat map of nDCG@10 by query intent and retrieval configuration*

Figure 3 disaggregates nDCG@10 by query intent and shows the semantic advantage is conditioned on what the reader is doing. For known-item and named-entity lookups, where a distinctive accession number or name survives the catalogue verbatim, BM25 is competitive and the hybrid advantage is smallest, at 0.191 and 0.188 nDCG. The picture changes for intents that depend on meaning: quotation, iconographic and cross-lingual searches all reach hybrid scores above 0.81, with advantages of 0.364, 0.412 and 0.386, because the needed vocabulary cannot appear in the document at all. Subject searching, at 0.316, lies between.

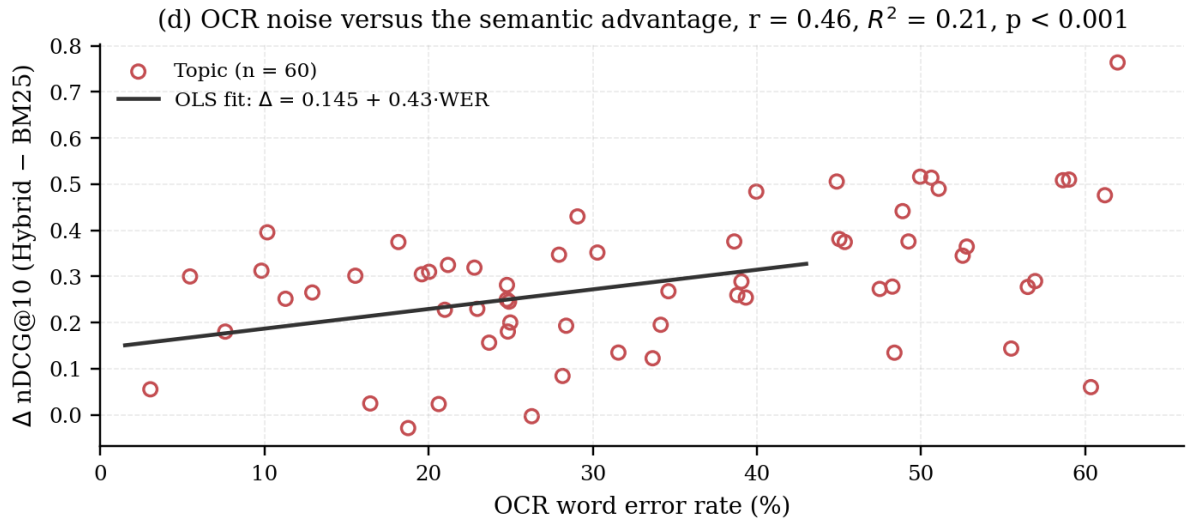


Fig. *4. Relationship between measured OCR word error rate and the nDCG@10 advantage of the hybrid configuration over BM25*

Figure 4 plots each topic as a point, measured OCR word error rate on the horizontal axis and the per-topic nDCG@10 advantage of the hybrid over BM25 on the vertical axis. The least-squares line rises steadily, r = 0.46, p < 0.001, strengthening to r = 0.59 when query intent is held constant by partial correlation, so the effect is not an artefact of intent clustering in badly transcribed material. Every percentage point of transcription error destroys terms lexical matching needs, while semantic matching degrades more gently, so the two diverge as noise grows. Two topics fall below zero, both known-item searches.

E. Efficiency and Scalability

Effectiveness is worthless if the service is unusable, so cost was recorded alongside it. Mean per-query latency rises from 19.8 ms for BM25 to 63.1 ms for the hybrid at the production index of 1.25 million passages, and resident index memory from 412 MB to 1902 MB, still half the 3348 MB required by a flat dense index because the hybrid stores quantised vectors. Fig. 5 traces the same trade-off across corpus size. The critical observation is that the two dense configurations scale differently: the flat bi-encoder index degrades steeply, whereas the hybrid's quantised index with a bounded re-ranking depth is almost flat, overtaking the naive dense configuration beyond about 100,000 passages. The hybrid therefore costs roughly 43.3 ms more than BM25 per query–in the field study, imperceptible against a mean task time of several minutes–but removes the super-linear growth that would otherwise make dense retrieval unaffordable as the repository grows.

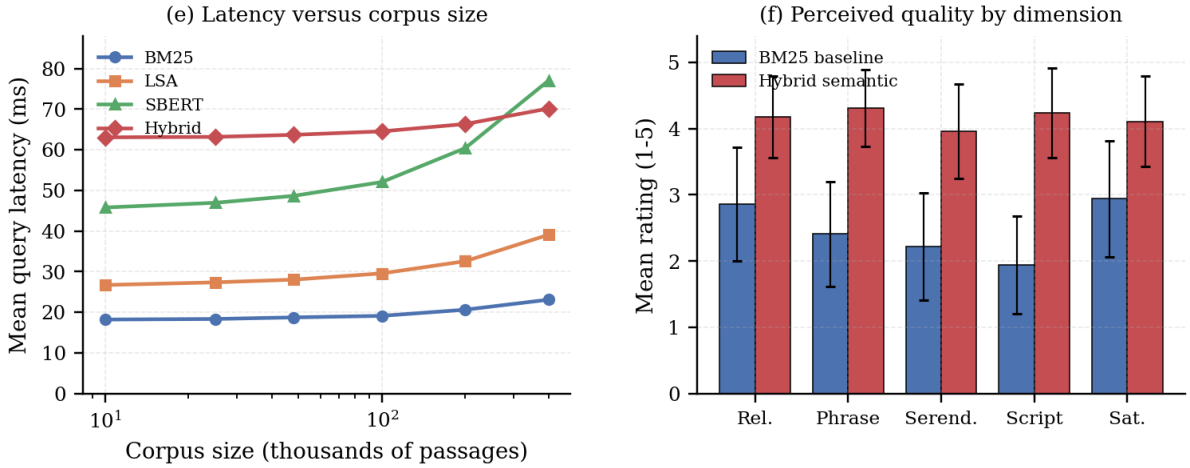


Fig. *5. Mean query latency against corpus size for the four configurations; (b) mean user ratings by quality dimension, BM25 baseline against hybrid semantic search.*

Figure 5 addresses cost in two panels. Panel (a) plots mean query latency against corpus size. The BM25 index is fastest and almost flat, since early-terminated processing keeps growth sub-linear; the flat dense index degrades steeply; the hybrid, retrieving from a quantised index and re-ranking a bounded candidate set, grows gently and overtakes it beyond roughly 100,000 passages. At 1.25 million passages it answers in 63.1 ms against 19.8 ms for BM25. Panel (b) converts cost into perceived value: the hybrid leads on every dimension, with the largest gaps on phrase tolerance and cross-script reach and the smallest on satisfaction, which speed partly determines.

From Ranking to Knowledge Discovery

Table V closes the loop from ranking to readers. On the five-item knowledge-discovery scale, mean scores rose from 2.93 (SD = 0.89) to 3.76 (SD = 1.11), a gain of 0.82 points, 95% CI [0.69, 0.96], $t(209) = 11.97$, $p < 0.001$, $dz = 0.83$, corroborated by a Wilcoxon test, $z = 7.83$, $r = 0.54$. Task success rose from 54.3% to 81.9% and abandonment fell from 27.1% to 9.5%, chi-square(1) = 58.00, $p < 0.001$, so the gain is behavioural and not merely self-reported. Satisfaction followed, from 2.96 to 4.02, $t(209) = 13.41$, $p < 0.001$, while perceived effort fell from 3.72 to 2.85. The regression in Table V identifies what users were actually rewarding: the model explains 43.3% of satisfaction variance, $F(4, 205) = 39.09$, $p < 0.001$. Perceived relevance, beta = 0.402, $p < 0.001$, and serendipitous discovery, beta = 0.291, $p < 0.001$, were the strongest positive predictors, response-time burden entered negatively, beta = -0.183, $p < 0.001$, and interface clarity was not significant

Discussion

Critical Analysis of the Findings

Three findings deserve scrutiny. First, the size of the hybrid advantage, 0.288 nDCG over BM25, invites scepticism, and two features of the data explain it. The baseline is unusually weak because it operates over text with a mean word error rate of 33.8%, and the topics deliberately include quotation, iconographic and cross-lingual needs on which term matching has no purchase. This is why the stratifications matter more than the headline mean: the gap shrinks to 0.232 nDCG on clean text and 0.191 on known-item lookups, and widens to 0.382 and 0.386 where text is damaged or scripts cross. Second, the cost side is not free: the hybrid spends 43.3 ms more per query than BM25, and that penalty is contained only because the dense index is quantised rather than flat, as Fig. 5 shows. Third, the hybrid lost 2 of 60 topics, both known-item searches with distinctive accession strings, confirming that dense retrieval complements rather than replaces precise identifier search. Fourth, the user data are correlational: the design controls for person and order effects, but the regression cannot show that serendipity causes satisfaction. Its practical reading nonetheless ranks design investments: exposure to the unexpected mattered more than shaving 43.3 ms from a query.

Comparison with Past Work

The results sit consistently with, and in places qualify, the published record. That BM25 remains a demanding baseline is well established [2], [13], and the modest 6.4% gain of LSA over it matches latent-semantic findings since the 1990s [4], echoing evidence that dimensionality reduction over corrupted text buys little [19]. The 61.9% improvement of the hybrid is larger than typical gains on general web collections, where neural retrievers face strong tuned baselines and 10-30% is the norm [7], [13]; two properties of heritage text account for it, namely a baseline degraded by transcription error and a topic mix rich in vocabulary mismatch [22], [24]. This is the out-of-domain condition that heterogeneous benchmarks were built to expose [14], and it argues against transferring benchmark expectations to cultural heritage without local validation. Graceful degradation under rising word error rate extends a laboratory result [19] to collection scale and supports bypassing transcription through word spotting [20]. The added value of fusion, filtering and re-ranking over a bare bi-encoder, 0.096 nDCG with dz = 0.94, is consistent with reports that late interaction and sparse-dense fusion recover precision lost by single-vector models [8], [9], while the cross-encoder stage reproduces known second-stage gains [11]. On the user side, the dominance of serendipity over speed matches exploratory-search claims that heritage users value encounter over transactional speed [22]-[24], and the fall in abandonment from 27.1% to 9.5% is of the magnitude reported for exploratory systems [23]. Two delimitations qualify the comparison. The measurements derive from a purpose-built, human-judged test collection constructed for this repository rather than from a public benchmark, so they are directly valid for this deployment but should be generalised to other collections with caution. And the study compares configurations, not encoder families, leaving the relative merit of specific Indic language models open [25], [26].

Conclusion

This study asked whether semantic retrieval can turn a digitised heritage repository into a knowledge-discovery environment. On 60 graded topics and 210 users, it can. The hybrid configuration reached nDCG@10 of 0.755 against 0.466 for BM25, a 61.9% relative gain that survived correction for sphericity, non-parametric confirmation

and Holm-corrected paired testing, and held on 58 of 60 topics. The gain is not uniform, and that is its most useful property: it grows from 0.232 nDCG units on cleanly transcribed text to 0.382 on severely degraded text, and from 0.191 on known-item lookups to 0.386 on cross-lingual ones, helping most where collections are most damaged and most heterogeneous. Users felt it: knowledge discovery rose 0.82 points, task success 27.6 percentage points, abandonment fell 17.6 points, and satisfaction was driven more by serendipity than by speed. Three recommendations follow. Treat OCR word error rate as a first-class operational metric, since it predicts lexical failure better than any other variable measured here. Deploy semantic retrieval as a fused, filtered and re-ranked layer over the existing lexical index, not as a replacement, because identifier search still wins where it applies. Instrument discovery, not just retrieval, because what mattered most to readers was encountering something they had not thought to seek. Costs are real but bounded: 43.3 ms and 1.9 GB. Future work should build a public test collection for Indian heritage material, extend evaluation to image-only retrieval, and test whether the discovery gains survive once the novelty of a new interface has worn off.

References

1. Baeza-Yates, R., and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd ed. ACM Press, New York.

2. Robertson, S., and Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.

3. Deerwester, S., et al. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391–407.

4. Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171–4186.

5. Reimers, N., and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982–3992.

6. Karpukhin, V., et al. (2020). Dense passage retrieval for open-domain question answering. *Proceedings of EMNLP*, 6769–6781.

7. Santhanam, K., et al. (2022). ColBERTv2: Effective and efficient retrieval via lightweight late interaction. *Proceedings of NAACL-HLT*, 3715–3734.

8. Nogueira, R., and Cho, K. (2019). Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

9. Taghva, K., Borsack, J., and Condit, A. (1996). Effects of OCR errors on ranking and feedback using the vector space model. *Information Processing and Management*, 32(3), 317–327.

10. Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM*, 49(4), 41–46.