Breaking the Top-k Assumption in Pseudo-Relevance Feedback: An Empirical Analysis of Relevance Distribution in BM25 Rankings

Khaled Albishre
Department of Computer Science, University College of Al Jamoum, Umm Al-Qura University, Al Jumum, Saudi Arabia
kmbishre@uqu.edu.sa (corresponding author)

Received: 21 March 2026 | Revised: 27 April 2026 and 16 May 2026 | Accepted: 17 May 2026

 DOI: https://doi.org/10.48084/etasr.18857

ABSTRACT

Pseudo-Relevance Feedback (PRF) has remained a cornerstone of unsupervised retrieval since Rocchio (1971), yet the foundational assumption that the top-k retrieved documents are the best available feedback has received limited direct empirical scrutiny, despite widespread adoption in both classical and neural approaches. This study analyzed BM25 retrieval on the TREC Deep Learning 2019 and 2020 test collections (97 queries, 20,646 graded relevance judgements) and found that 79% of relevant documents fall outside the top-10, with a mean rank of 37.3. An oracle selection strategy achieved 0.324 higher feedback precision at $k = 10$, defined as the proportion of graded-relevant documents within the k selected for expansion, with a large effect size that is consistent across all tested values of k and both test collections. LLM-based analysis of 61 extreme cases identified vocabulary gap as the dominant failure mode in 96.7% of cases, driven primarily by implicit relevance (35.6%) and hypernym-hyponym mismatch (27.1%). These findings establish that document selection, rather than term weighting, is the primary lever for PRF improvement and identify the vocabulary gap as the principal target for next-generation methods. The results demonstrate that improving feedback-document selection represents a largely unexplored avenue for PRF advancement.

Keywords-pseudo-relevance feedback; BM25; TREC deep learning; vocabulary gap; oracle experiment; query expansion

I. INTRODUCTION

Pseudo-Relevance Feedback (PRF) is among the most effective unsupervised techniques in information retrieval, and it remains central to addressing the persistent query-document vocabulary mismatch. The scope of this mismatch has been documented in detail in [1], in a survey of scholarly retrieval systems that framed vocabulary divergence between user intent and document content and defined the limitation of lexical search and the primary motivation for downstream expansion techniques. Since Rocchio's original formulation in the vector space model [2], PRF methods have consistently improved retrieval effectiveness across diverse benchmarks [3, 4]. Representative approaches include RM3 [5, 6], divergence-based models [7, 8], and the comparative term-selection study in [9]. These methods assume that the top-ranked documents from an initial retrieval are relevant, extract informative terms from them, and use those terms to expand or reformulate the original query. In every case, the top-k documents serve as the exclusive feedback source.

Recent work has extended PRF into the dense and generative-retrieval era. ANCE-PRF [10] augments dense query representations using top-k passages, ColBERT-PRF [11] applies feedback within the ColBERT late-interaction framework, and in [12], PRF generalizes to multiple-representation dense retrieval. On the generative side, GPRF [13] uses reinforcement learning to rewrite queries from retrieved feedback documents, LLM-VPRF [14] extends vector-based PRF to LLM-based dense retrievers, and PromptPRF [15] uses LLM-extracted features from top-ranked documents to refine dense query representations. Dense-chain approaches that refine feedback sequentially [16] represent a step toward addressing document selection in neural PRF. The architectural innovations are varied, but the input is not: all of these methods feed on the same top-k documents that Rocchio would have used.

The foundation of this mechanism is a single, largely unexamined assumption: that the top-k documents retrieved by the initial ranker are the best available source of relevance feedback. A document at rank 1 is presumed to be more useful as feedback than one at rank 50. Prior work on selective PRF [17], feedback filtering [3], and query drift has examined aspects of feedback quality. Yet the top-k assumption itself—that rank position is a reliable proxy for feedback utility—has received limited direct empirical scrutiny. In [17], a selective method skips feedback when the initial retrieval appears

unreliable, but when feedback proceeds, it uses the standard top-k. Automatic training-set selection for feedback documents has been explored in short-text retrieval [18], although the fundamental top-k assumption was not questioned. In [3], PRF pitfalls were surveyed, noting that noise in the top-k can degrade performance, while in [19], a systematic study was presented but did not focus on the selection mechanism itself.

The reason to doubt this assumption is straightforward. First-stage retrievers such as BM25 [20] rank documents by lexical similarity, not by relevance. A document may be highly relevant to a query while employing entirely different vocabulary, a phenomenon termed the vocabulary gap. If BM25 rank is a poor proxy for actual relevance, the top-k assumption breaks down, and PRF methods are constructing feedback from the wrong documents.

A parallel body of work addresses the vocabulary gap from the query side. HyDE [21] generates a hypothetical document from the query using a language model and then retrieves it against the generated text. Query2Doc [22] augments queries with LLM-generated passages to bridge lexical mismatches, and ThinkQE [23] iteratively refines query expansions through a thinking-based process combined with corpus-interaction feedback. The scholarly-search literature offers an equally important and arguably more interpretable line of evidence: In [24], it was shown that classical query expansion, when combined with citation-network analysis, yields measurable gains in relevance—a result that demonstrates the durability of term-side enrichment as a complement to neural and document-side methods, particularly in domains where document graphs carry strong semantic signals. Vocabulary mismatch has similarly been identified as a core obstacle in short-text retrieval environments, motivating unsupervised approaches that bridge the lexical gap between queries and documents [25]. Although these query-side methods target the vocabulary gap upstream, this work reveals the same problem from the other side: relevant documents that BM25 ranks poorly due to vocabulary mismatch are excluded from PRF feedback, regardless of how well the query has been expanded. The two perspectives are complementary.

To our knowledge, no prior work has measured the empirical distribution of relevant documents across BM25 rankings, quantified the feedback precision gap between top-k and relevance-optimal document selection, or provided a taxonomy of why BM25 systematically misplaces relevant documents in the context of passage retrieval on standard TREC benchmarks. This paper addresses these three questions directly.

Using BM25 retrieval on the TREC Deep Learning 2019 [26] and 2020 [27] test collections (97 queries, 20,646 graded relevance judgements), this study makes four contributions. First, it quantifies the distribution of relevant documents across BM25 top-100 rankings; the result is stark: 79% of relevant documents fall outside the top-10. Second, an oracle experiment reveals a feedback precision gap of +0.324 at $k = 10$ ($p < 0.001$, Cohen's d = 1.173), where feedback precision is the proportion of graded-relevant documents within the k selected for expansion. Third, an LLM-based taxonomy of six vocabulary-gap subtypes from 61 extreme misranking cases is constructed, revealing implicit relevance (35.6%) and hypernym-hyponym mismatch (27.1%) as dominant failure modes. Fourth, this study shows that the BM25 score correlates weakly with relevance (Spearman $\rho = 0.203$), accounting for only about 4% of the variance in human judgements.

II. METHODOLOGY

A. Datasets

All experiments operate at the passage-retrieval level. The TREC Deep Learning 2019 [26] and 2020 [27] test collections were used, which were built on the MS MARCO v1 passage corpus [28] (approximately 8.8 million passages). Table I summarizes the dataset statistics. TREC DL 2019 provides 43 queries with 9,260 relevance judgements; DL 2020 provides 54 queries with 11,386 judgements. Combined, the two tracks provide 97 queries and 20,646 graded judgements on a four-point scale (0–3).

TABLE I. DATASET STATISTICS

	DL 2019	**DL 2020**	**Combined**
Queries	43	54	97
Judgements	9,260	11,386	20,646
Relevance scale	0–3	0–3	0–3

B. Retrieval Setup

BM25 [20] was employed via Pyserini [29] with default parameters ($k1 = 0.9, b = 0.4$) on the pre-built msmarco-v1-passage index, retrieving the top-100 documents per query. BM25 is the natural choice: it remains the dominant first-stage retriever in production and serves as the default initial ranker for virtually all PRF methods [2, 11, 13].

C. Analysis Framework

1) Relevance Distribution

The rank position, relevance grade, and rank bucket (top-10, 11-20, 21-50, 51-100) were recorded for every relevant document ($grade \geq 1$) in the top-100. In addition, coverage curves were computed at $k \in \{1, 3, 5, 10, 20, 30, 50, 100\}$.

2) Oracle PRF Experiment

Standard PRF (select top-k by BM25 rank) was compared against oracle PRF (select the k most relevant documents from the top-100 pool, by $qrels$ grade). Formally, let $C_q = top - 100(BM25, q)$ denote the candidate pool retrieved for query q. Standard PRF returns the first k documents of C_q ordered by BM25 score. Oracle PRF returns the k documents of C_q with the highest TREC relevance grades, with ties broken by BM25 rank. Both methods draw from the same candidate set; they differ only in the selection criterion. The oracle measures the ceiling of document-selection quality attainable within what the first-stage retriever already exposes—not the ceiling of an improved retriever. Because it uses ground-truth labels unavailable at query time, the oracle represents an upper bound, not a deployable target. The metric is Feedback Precision@ k, defined as the fraction of the k selected documents with $TREC\ grade \geq 1$. Tests were conducted for $k \in \{3, 5, 10, 20\}$.

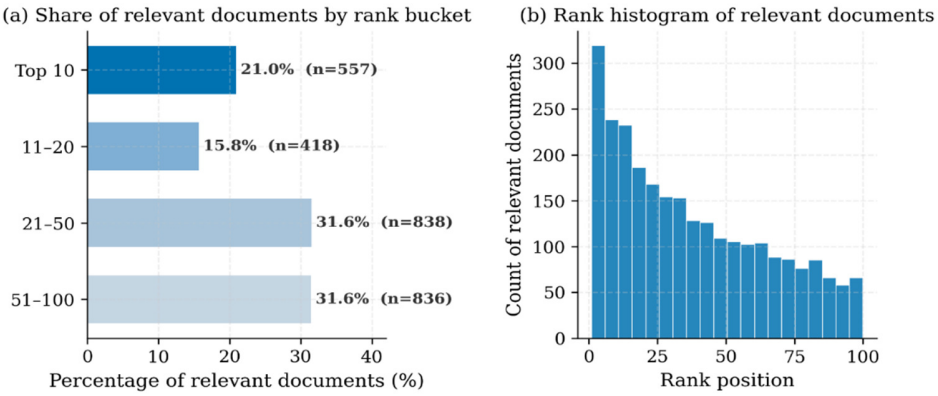

Fig. 1. Relevance distribution across BM25 rank buckets (left) and histogram of rank positions (right).

3) Statistical Testing

Differences were assessed using the Wilcoxon signed-rank test (two-sided, $\alpha = 0.05$) with Cohen's d on paired per-query differences.

D. Vocabulary Gap Analysis

Query term coverage was measured for each (query, document) pair: $Coverage = |Q \cap D|/|Q|$, where Q and D denote the sets of non-stopword terms in the query and document, respectively. The Spearman correlation between coverage and rank among relevant documents was computed.

E. LLM-Based Taxonomy

This study identified 61 extreme cases: relevant documents with coverage below 0.3, ranked 51–100. For each case, Mistral-7B-Instruct [30] received the query, the document, and the following instruction: "Given the query Q and document D, determine whether D is relevant to Q. If relevant, identify the principal lexical or semantic mismatch that could cause a lexical retriever to rank D poorly, and explain briefly." The author manually reviewed all 61 outputs, correcting the model's judgment in 2 cases. The 59 vocabulary-gap cases were coded into subtypes following an open-coding procedure, inductively deriving categories rather than imposing a predefined taxonomy. Coding was performed in two passes separated by 48 hours to verify internal consistency; passes agreed on 58 of 59 cases (98.3%). Only a single human annotator performed the coding; this limitation is stated explicitly in Section V.

III. RESULTS

Throughout this section, statistical tests use the Wilcoxon signed-rank test ($\alpha = 0.05$), and effect sizes are reported as Cohen's d.

A. Relevance Distribution in BM25 Rankings

Across 97 queries, BM25 retrieved 2,649 relevant documents ($grade \geq 1$) within the top-100. Figure 1 shows how these documents distribute across rank positions. Only 21.0% of relevant documents appear in the top-10, the feedback window that PRF methods rely upon. The remaining 79.0% are spread across ranks 11–100. The two deepest buckets (21–50 and 51–100) each contain 31.6% of relevant documents, more than the top-10 itself. The mean rank of a relevant document is 37.3.

The pattern holds across queries. Over a quarter (25.8%) of queries have a non-relevant document at rank 1—the single result PRF trusts most. Both test collections tell the same story: DL 2019 shows 80.6% outside the top-10 (mean rank 38.7); DL 2020 shows 77.2% (mean rank 35.9).

Grade-3 documents rank somewhat higher on average (mean rank 32.1) than grade-1 documents (mean rank 38.5); yet 71.4% of them still fall outside the top-10. Higher relevance does not protect a document from BM25's lexical bias.

B. Top-k Coverage Analysis

To quantify how much relevant evidence reaches PRF at each retrieval depth, the percentage of all relevant documents captured within the top-k pool were computed for $k \in \{1, 3, 5, 10, 20, 30, 50, 100\}$. Figure 2 plots these coverage curves separately for all relevant documents ($grade \geq 1$) and for the most relevant subset (grade 3).

At $k = 10$, BM25 captures only 7.2% of all relevant documents. At $k = 100$, still only 34.4%. The implication is direct: standard PRF does not operate on a noisy version of the relevant signal—it operates on a fundamentally incomplete one, missing over 92% of the relevant evidence at the typical feedback depth.

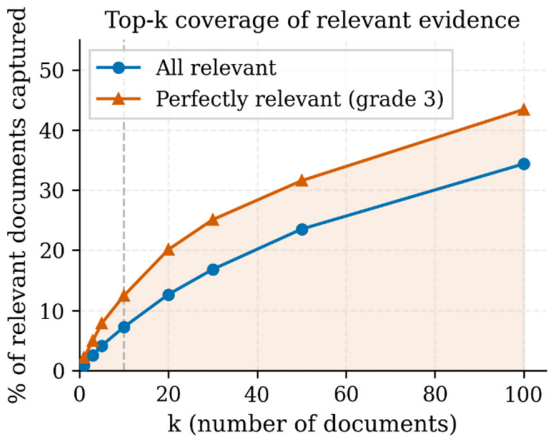

Fig. 2. Coverage curve: percentage of relevant documents captured at each retrieval depth k.

C. Oracle PRF vs. Standard PRF

To measure the practical cost of restricting feedback to the top-k by rank, an oracle was constructed that selects the k most relevant documents from the full top-100 pool using ground-truth relevance grades, irrespective of rank position. This oracle represents an upper bound on feedback precision achievable without modifying the retrieval model. Table II and Figure 3 present the results.

TABLE II. ORACLE VS. STANDARD PRF FEEDBACK PRECISION

k	Standard	Oracle	Gap	p-value	Cohen's d
3	0.670	0.983	+0.313	2.7×10^{-10}	0.886
5	0.658	0.959	+0.301	1.9×10^{-11}	0.940
10	0.574	0.898	+0.324	4.9×10^{-14}	1.173
20	0.503	0.763	+0.261	7.8×10^{-16}	1.292

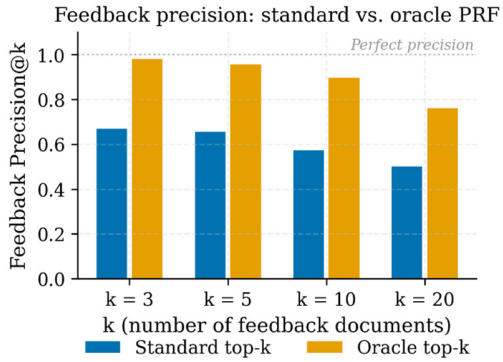

Fig. 3. Oracle PRF vs. standard PRF feedback precision across k values.

At $k = 10$, standard PRF achieves feedback precision of 0.574, meaning 42.6% of its feedback documents are non-relevant. The oracle reaches 0.898. The gap of +0.324 is highly significant (Wilcoxon $p = 4.9\times10^{-14}$, Cohen's $d = 1.173$). The gap is consistent across all tested values of k: +0.313 at $k = 3$ (d =0.886), +0.301 at $k = 5$ (d =0.940), and +0.261 at $k = 20$ (d =1.292). Per-collection analysis confirms robustness: the gap reaches +0.319 on DL 2019 (p =5.1×10^{-7}) and +0.328 on DL 2020 (p =1.6×10^{-8}), indicating that the finding is not an artifact of any single test collection.

A feedback selection method that approximates oracle behavior, even imperfectly, can recover a substantial portion of this precision gap. No change to the underlying retrieval model is required. The oracle does not require a better retriever—the relevant documents are already present in the top-100. What is missing is a selection criterion that ranks them above the non-relevant ones. Closing even half the gap would raise standard PRF feedback precision from 0.574 to approximately 0.736 at k =10. This gap exceeds typical gains reported from term-weighting refinements in the PRF literature [5, 7, 9].

D. Query Difficulty Analysis

A natural expectation is that relevant-document scattering is a property of difficult queries—those where BM25 performs poorly overall. Figure 4 tests this directly, and the result is unambiguous: the problem is not confined to hard queries.

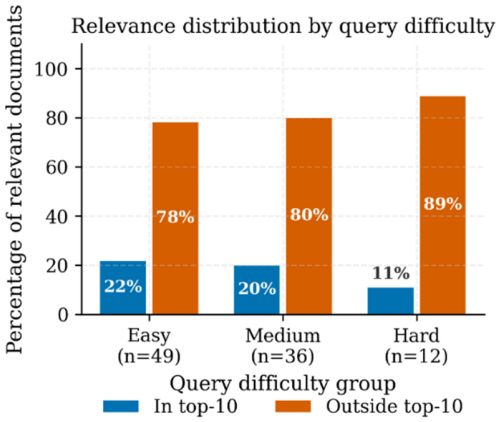

Fig. 4. Relevance distribution by query difficulty group.

- Easy queries (n =49): 78% of relevant documents outside the top-10, mean rank 31.
- Medium (n =36): 80%, mean rank 32.
- Hard (n =12): 89%, mean rank 34.
- Even when BM25 performs well, the vast majority of relevant evidence remains beyond the PRF window. The top-k assumption fails across the full difficulty spectrum.

What stands out here is not just that hard queries scatter more, but how little the mean rank changes across bands (31, 32, 34). If lexical mismatch was mainly a symptom of poorly worded or ambiguous queries, the gap between easy and hard queries should be much wider. Instead, it looks as if vocabulary divergence between a query and its truly relevant documents is something that happens regardless of how well the query is phrased. The practical reading of this is that PRF improvements measured only on easy queries should not be expected to carry over to harder ones at the same magnitude, and methods that target hard queries by trying to improve the query itself are unlikely to fix the problem.

The hard-query band is also where the picture is most extreme. With 89% of relevant documents outside the top-10, the standard feedback window for these queries contains almost nothing useful to learn from. For this subset, the top-k assumption is not just suboptimal in the sense of leaving performance on the table; it is selecting feedback from a region that is mostly empty of signal. These are the queries where a better selection rule, even a simple one, should make the largest difference in practice.

E. BM25 Score-Relevance Correlation

The distribution documented in Section III.A raises a fundamental question: to what extent does the BM25 score reflect actual relevance? A weak correlation between BM25 score and human relevance judgement would mean that rank position is an unreliable proxy for feedback quality, and if rank is unreliable, the top-k assumption loses its only justification.

Across 9,700 (query, document) pairs, the Spearman correlation between BM25 score and relevance grade is $\rho = 0.203$ ($p < 0.001$). The correlation is statistically

significant but practically weak: BM25 score accounts for only about 4% of the variance in relevance ($\rho^2 = 0.041$). Figure 5 makes this concrete: the score distributions for grades 0 through 3 overlap extensively.

This is interpreted as follows: BM25's lexical matching captures enough of relevance to populate the top-100 with relevant documents, but not enough to sort them correctly within that pool. This weak discrimination is the root cause of the scattering documented in Section III.A and the feedback quality gap in Section III.C.

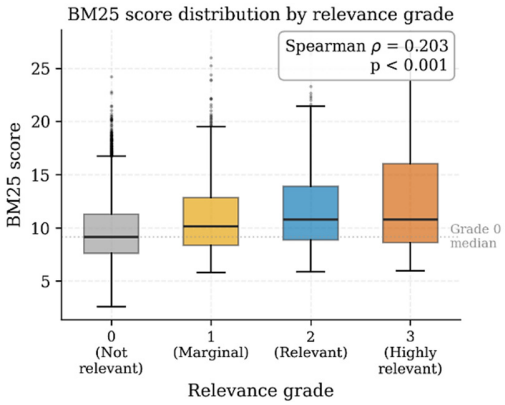

Fig. 5. BM25 score distribution by relevance grade.

IV. WHY DOES BM25 FAIL?

The preceding analysis establishes the phenomenon: relevant documents are distributed across the full top-100 rather than concentrated at the top, and BM25 scores correlate weakly with human relevance judgements ($\rho = 0.203$). The remaining question is why BM25 fails to surface relevant documents at higher ranks. This section investigates the mechanism.

A. *Vocabulary Gap Analysis*

The hypothesis is straightforward: relevant documents ranked deeply in the list employ vocabulary that differs from the query terms. Query term coverage was computed for all 2,649 relevant documents in the pool. Coverage is defined as the proportion of non-stopword query terms that also appear in the document—a measure of surface-level lexical overlap that BM25 implicitly rewards. If this hypothesis is correct, relevant documents at deeper ranks should exhibit systematically lower coverage than those at higher ranks.

The relation between coverage and rank was examined from three complementary angles. First, coverage was plotted against rank position for every relevant document, separated by relevance grade, to inspect whether the trend is uniform across grades or driven by a particular subset. Second, coverage was aggregated into rank buckets (top-10, 11–20, 21–50, 51–100) to quantify the magnitude of the drop. Third, coverage was examined by relevance grade to verify that the pattern is not an artifact of grade-1 documents alone. Figure 6 presents the three views. The data confirm the hypothesis clearly. Relevant documents in the top-10 have a mean coverage of 0.755. Coverage drops steadily: 0.672 at ranks 11–20, 0.633 at 21–50, 0.604 at 51–100. The Spearman correlation between coverage and rank is $\rho = -0.234$ ($p < 10^{-33}$). Relevant documents with low lexical overlap are systematically pushed down the ranking. The coverage difference of 0.151 between the top-10 and ranks 51–100 translates directly into lower BM25 scores. Documents that discuss the query topic using synonyms, hypernyms, or domain-specific terms are ranked below non-relevant ones that share more surface terms with the query.

B. *Document Length: A Null Result*

Length bias was also considered. Perhaps most surprisingly, this explanation finds no support. The Spearman correlation between document length and rank gap is $\rho = 0.015$ ($p = 0.43$). MS MARCO passages are uniformly short (50–150 words), offering too little variation for this mechanism to operate. The vocabulary gap alone accounts for the observed displacement.

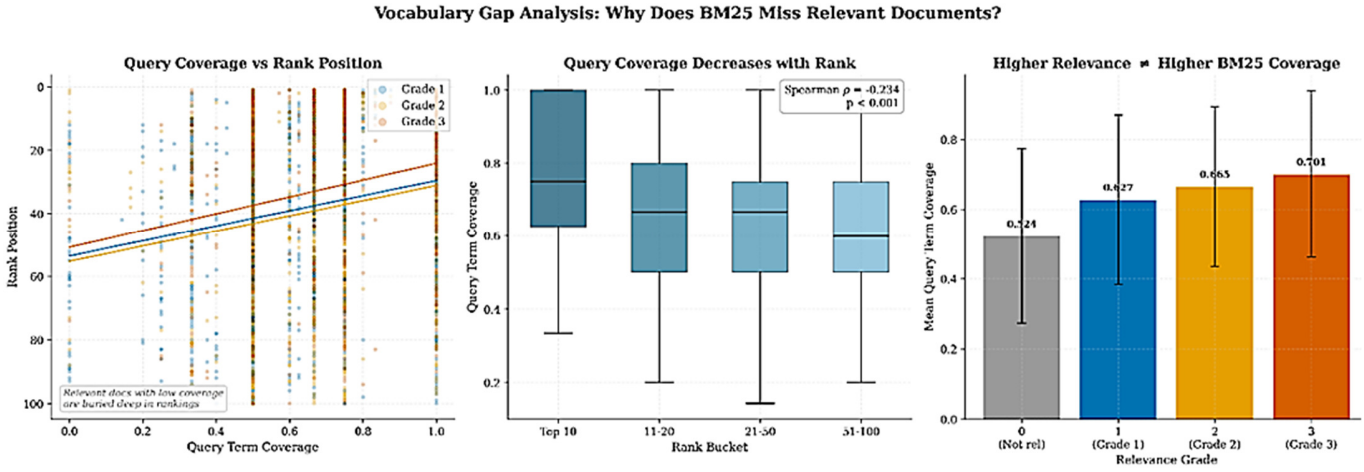

Fig. 6. Vocabulary-gap analysis: (a) coverage vs. rank scatter plot, (b) coverage by rank bucket, (c) coverage by relevance grade.

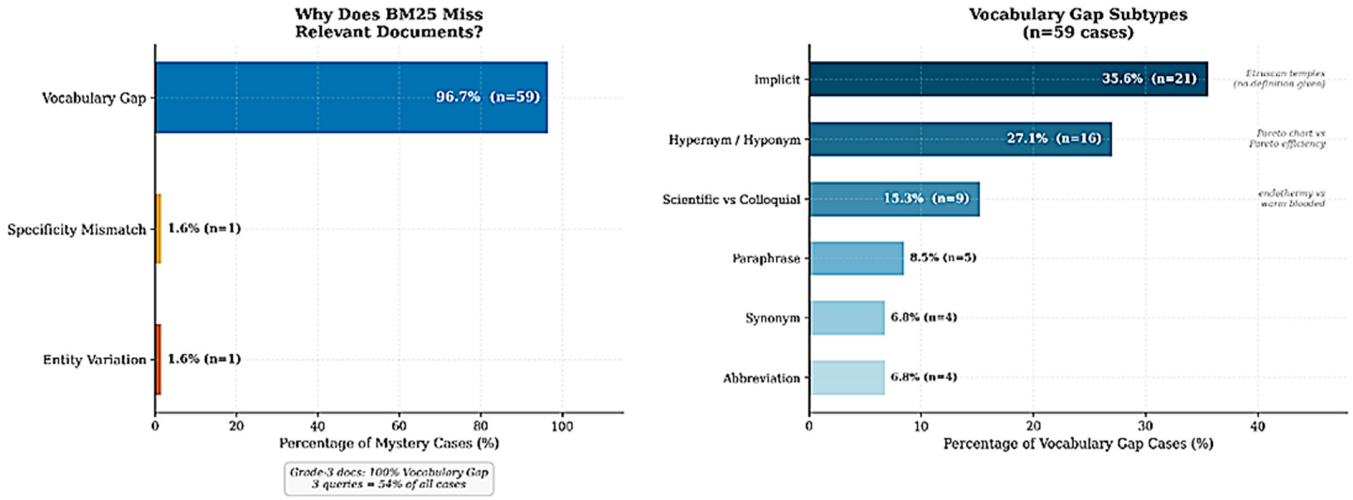

Fig. 7. LLM taxonomy: main failure categories (left) and vocabulary-gap subtypes (right).

C. LLM Taxonomy of the Vocabulary Gap

61 extreme cases were examined: relevant documents with coverage below 0.3, ranked 51–100. These documents share almost no terms with the query, yet TREC assessors judged them relevant. Mistral-7B-Instruct [30] was prompted to classify and explain each case. The results, summarized in Figure 7, reveal a highly consistent pattern: 59 of 61 cases (96.7%) involve vocabulary gap as the primary failure mode. The remaining two involved specificity mismatch and entity variation. Qualitative coding yielded six subtypes: implicit relevance (35.6%, n =21), where the document discusses the query topic without using the query terms; hypernym-hyponym mismatch (27.1%, n = 16), where query and document use related terms at different specificity levels; scientific versus colloquial terminology (15.3%, n = 9); paraphrase (8.5%, n =5); synonym (6.8%, n = 4); and abbreviation (6.8%, n = 4).

Two patterns in the data merit particular attention. All six grade-3 documents among the low-coverage, deep-ranked cases were vocabulary gap—the most valuable documents BM25 misses are uniformly lost to lexical mismatch. And the cases concentrate heavily: three queries account for 33 of 61 cases (54%), indicating that topics with rich alternative vocabularies are disproportionately vulnerable.

V. DISCUSSION

A. Implications for PRF Design

The finding that 79% of relevant documents sit outside the top-10 feedback window has a clear practical consequence. The oracle experiment shows that this restriction costs +0.261 to +0.324 in feedback precision, with effect sizes consistently above 0.88. The gap is not marginal. A broader conclusion is drawn: feedback document selection deserves at least as much research attention as the feedback model. The community has devoted decades to improving term-weighting schemes, while the input to these models has remained the same top- k documents. The results of this study indicate that improving the input may matter as much as improving the processing.

This analysis points to three concrete signals, each a practical approximation of the oracle principle. Dense embedding cosine similarity would favor topically related documents regardless of lexical overlap. Query term coverage thresholds offer a simpler alternative: documents with coverage below 0.3 are vocabulary-gap cases 96.7% of the time. A lightweight cross-encoder applied to the top-100 would approximate oracle selection more closely. These signals could be deployed individually or combined.

B. Connection to Reranking

Neural rerankers, such as monoT5 [31], improve retrieval output—the list the user sees. This work motivates improving PRF input—the documents that inform expansion. A full reranker before PRF would help, but may be too expensive for latency-sensitive applications. The coverage and similarity signals proposed above offer lighter alternatives.

C. Generalization to Dense and Hybrid Retrievers

The findings of this study are established for BM25 first-stage retrieval, which remains the default initial ranker for nearly all classical and dense-retrieval PRF systems, including ANCE-PRF [10], ColBERT-PRF [11], GPRF [13], and PromptPRF [15]. The vocabulary-gap mechanism identified in this study is intrinsic to lexical matching, and dense retrievers—designed in part to bridge semantic gaps via learned representations—would plausibly exhibit a weaker distributional problem. Existing evidence suggests, however, that dense retrievers reduce rather than eliminate lexical bias, particularly on out-of-domain queries. Quantifying the oracle gap under dense (e.g., DPR, ColBERT, SPLADE) and hybrid (lexical+dense fusion) retrievers is a direct extension of this work and is left to future research.

D. Beyond Binary Relevance

A complementary perspective on feedback-document quality is offered by multi-criteria approaches to relevance. In [32], it was argued that the usefulness of a retrieved document in academic search depends not only on topical match but also on factors such as publication recency, citation impact, and

domain-specific signals—a multi-objective view of relevance that has direct implications for what counts as a high-quality feedback document. This framework suggests that future definitions of feedback-document quality, once enriched beyond the graded relevance labels used here, could combine topical fit with these additional dimensions to produce a more nuanced selection criterion than either rank position or topical relevance alone can provide.

E. Limitations

This analysis covers TREC DL 2019 and 2020, both based on MS MARCO web queries. The pattern may differ on news (TREC Robust), multilingual (CLEF), or domain-specific corpora. The 97-query set is standard but modest in size. The oracle uses ground-truth relevance labels unavailable at query time; it therefore represents an upper bound on what document selection alone can achieve, not a directly deployable target. The LLM taxonomy was classified by a single annotator with manual verification; no inter-annotator agreement metric was computed, and subtype proportions may not generalize to other collections or query types.

VI. CONCLUSION

PRF has remained one of the most widely used unsupervised techniques in information retrieval for more than fifty years, and every variant—from Rocchio (1971) through RM3 and divergence-based models to recent dense and LLM-based formulations—has rested on the same operational decision: the top-k documents returned by the initial ranker are taken as the best available feedback. This paper has subjected that decision to direct empirical scrutiny on standard TREC deep learning passage-retrieval collections, and the evidence accumulated across four orthogonal analyses converges on a single conclusion: the top-k assumption is the dominant point of failure in the contemporary PRF pipeline, and the failure is one of document selection rather than of term weighting or model architecture.

The knowledge gap addressed by this study lies precisely at the interface between first-stage retrieval and feedback construction. Decades of PRF research have refined how feedback terms are weighted, mixed, and re-injected into the query, but the question of which documents should enter that pipeline in the first place has been answered almost exclusively by the rank position assigned by the lexical retriever. By treating that selection step as an empirical variable rather than a fixed assumption, it was shown that the headline number—79% of relevant documents lying outside the top-10 feedback window, with a mean rank of 37.3—is neither marginal nor restricted to pathological queries; it persists across easy, medium, and hard query bands and across both DL 2019 and DL 2020 collections.

The four primary quantitative results of the study can be stated compactly. First, 79% of relevant documents fall outside the top-10 feedback window across 97 queries and 20,646 graded judgements. Second, an oracle that selects the k most relevant documents from the same BM25 top-100 candidate pool achieves +0.324 higher feedback precision at $k = 10$ (Wilcoxon $p = 4.9\times10^{-14}$, Cohen's $d = 1.173$), with the effect persisting across $k \in \{3, 5, 10, 20\}$ and across both test collections. Third, the Spearman correlation between BM25 score and human relevance grade is only $\rho = 0.203$, explaining roughly 4% of the variance in human judgements—sufficient to populate the top-100 with relevant documents but not to order them correctly within that pool. Fourth, vocabulary gap accounts for 96.7% of the extreme misranking cases, with implicit relevance (35.6%) and hypernym-hyponym mismatch (27.1%) emerging as the dominant failure subtypes.

The novelty of these results, relative to prior PRF work, lies in the combination of three elements that have not previously been brought together on standard TREC benchmarks: a fine-grained distributional analysis of where relevant documents actually sit in BM25 rankings, an oracle experiment that isolates the document-selection contribution from term-weighting effects, and a vocabulary-gap taxonomy grounded in LLM-generated rationales over extreme cases. Earlier studies on selective PRF, feedback filtering, and PRF pitfalls have approached related questions from adjacent angles, but none has quantified the empirical upper bound that document selection alone can contribute to feedback quality under a fixed first-stage retriever.

The contribution to the broader literature is correspondingly clear. The +0.324 oracle gap exceeds the typical improvements reported from term-weighting refinements in the PRF literature [5, 7, 9], which suggests that the marginal research return on better document selection is at least as large as that on better feedback models. The oracle gap is not a theoretical ceiling: it is the measurable distance between what PRF currently consumes and what is already present in the retrieved pool. No better retriever is required; the relevant documents are already there. What is missing is a selection criterion sensitive to vocabulary mismatch rather than to rank position.

Three concrete directions for future research follow directly from these findings. First, the three deployable signals identified in Section V—dense embedding cosine similarity, query-term coverage thresholds, and lightweight cross-encoder rescoring of the top-100—should be evaluated individually and in combination as drop-in replacements for the top-k selection rule in classical and neural PRF pipelines. Second, the oracle experiment should be replicated under dense (DPR, ColBERT, SPLADE) and hybrid first-stage retrievers to determine whether the vocabulary-gap mechanism and the magnitude of the oracle gap persist when lexical bias is partially mitigated upstream. Third, the LLM-based taxonomy should be extended beyond the single-annotator setting used here to multi-annotator validation on broader benchmarks (TREC Robust, BEIR, CLEF), allowing the subtype proportions reported here to be confirmed or refined across collection types and query distributions. Taken together, these directions outline an empirical research agenda for next-generation PRF systems in which feedback-document selection, rather than term weighting alone, is treated as a first-class design problem.

DECLARATION OF CONFLICTING INTERESTS

The author declares no conflicting or financial interests that could have affected the results of this study.

ACKNOWLEDGMENT

The author extends their appreciation to Umm Al-Qura University, Saudi Arabia, for funding this research work through grant number 26UQU4320004GSSR02.

DATA AVAILABILITY

The datasets used in this study are publicly available. The TREC Deep Learning 2019 and 2020 qrels are accessible at [26, 27], and the MS MARCO v1 passage index is available via Pyserini [29].

REFERENCES

[1] S. Khalid, S. Khusro, I. Ullah, and G. Dawson-Amoah, "On The Current State of Scholarly Retrieval Systems," *Engineering, Technology & Applied Science Research*, vol. 9, no. 1, pp. 3863–3870, Feb. 2019, https://doi.org/10.48084/etasr.2448.

[2] J. J. Rocchio Jr, "Relevance feedback in information retrieval," *The SMART retrieval system: experiments in automatic document processing*, pp. 313–323, 1971.

[3] H. Li, A. Mourad, S. Zhuang, B. Koopman, and G. Zuccon, "Pseudo Relevance Feedback with Deep Language Models and Dense Retrievers: Successes and Pitfalls," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–40, July 2023, https://doi.org/10.1145/3570724.

[4] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models." arXiv, Oct. 21, 2021, https://doi.org/10.48550/arXiv.2104.08663.

[5] N. Abdul-Jaleel *et al.*, "UMass at TREC 2004: Novelty and HARD:," Defense Technical Information Center, Jan. 2004. https://doi.org/10.21236/ADA460118.

[6] V. Lavrenko and W. B. Croft, "Relevance-Based Language Models," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 260–267, Aug. 2017, https://doi.org/10.1145/3130348.3130376.

[7] G. Amati, "Probability models for information retrieval based on divergence from randomness," Ph.D. dissertation, University of Glasgow, UK, 2003.

[8] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proceedings of the tenth international conference on Information and knowledge management*, July 2001, pp. 403–410, https://doi.org/10.1145/502585.502654.

[9] Y. Lv and C. Zhai, "A comparative study of methods for estimating query language models with pseudo feedback," in *Proceedings of the 18th ACM conference on Information and knowledge management*, Aug. 2009, pp. 1895–1898, https://doi.org/10.1145/1645953.1646259.

[10] H. Yu, C. Xiong, and J. Callan, "Improving Query Representations for Dense Retrieval with Pseudo Relevance Feedback," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Oct. 2021, pp. 3592–3596, https://doi.org/10.1145/3459637.3482124.

[11] X. Wang, C. MacDonald, N. Tonellotto, and I. Ounis, "ColBERT-PRF: Semantic Pseudo-Relevance Feedback for Dense Passage and Document Retrieval," *ACM Transactions on the Web*, vol. 17, no. 1, pp. 1–39, Feb. 2023, https://doi.org/10.1145/3572405.

[12] X. Wang, C. Macdonald, N. Tonellotto, and I. Ounis, "Pseudo-Relevance Feedback for Multiple Representation Dense Retrieval," in *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, July 2021, pp. 297–306, https://doi.org/10.1145/3471158.3472250.

[13] Y. Tu *et al.*, "Generalized Pseudo-Relevance Feedback." arXiv, Oct. 29, 2025, https://doi.org/10.48550/arXiv.2510.25488.

[14] H. Li, S. Zhuang, B. Koopman, and G. Zuccon, "LLM-VPRF: Large Language Model Based Vector Pseudo Relevance Feedback." arXiv, Apr. 02, 2025, https://doi.org/10.48550/arXiv.2504.01448.

[15] H. Li, X. Wang, B. Koopman, and G. Zuccon, "Pseudo Relevance Feedback is Enough to Close the Gap Between Small and Large Dense Retrieval Models." arXiv, June 06, 2025, https://doi.org/10.48550/arXiv.2503.14887.

[16] K. Albishre, "Anchored dense-chain pseudo-relevance feedback: sequential state refinement for neural retrieval," *Journal of King Saud University Computer and Information Sciences*, Mar. 2026, https://doi.org/10.1007/s44443-026-00604-x.

[17] S. Datta, D. Ganguly, S. MacAvaney, and D. Greene, "A Deep Learning Approach for Selective Relevance Feedback," in *Advances in Information Retrieval*, vol. 14609, N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, and I. Ounis, Eds. Cham: Springer Nature Switzerland, 2024, pp. 189–204.

[18] K. Albishre, Y. Li, and Y. Xu, "Query-Based Automatic Training Set Selection for Microblog Retrieval," in *Advances in Knowledge Discovery and Data Mining*, 2018, pp. 325–336, https://doi.org/10.1007/978-3-319-93037-4_26.

[19] N. Jedidi and J. Lin, "A Systematic Study of Pseudo-Relevance Feedback with LLMs." arXiv, 2026, https://doi.org/10.48550/ARXIV.2603.11008.

[20] S. Robertson and H. Zaragoza, *The Probabilistic Relevance Framework: BM25 and Beyond*. Now Publishers Inc, 2009.

[21] L. Gao, X. Ma, J. Lin, and J. Callan, "Precise Zero-Shot Dense Retrieval without Relevance Labels," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Apr. 2023, pp. 1762–1777, https://doi.org/10.18653/v1/2023.acl-long.99.

[22] L. Wang, N. Yang, and F. Wei, "Query2doc: Query Expansion with Large Language Models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Sept. 2023, pp. 9414–9423, https://doi.org/10.18653/v1/2023.emnlp-main.585.

[23] Y. Lei, T. Shen, and A. Yates, "ThinkQE: Query Expansion via an Evolving Thinking Process." arXiv, Mar. 09, 2026, https://doi.org/10.48550/arXiv.2506.09260.

[24] S. Khalid and S. Wu, "Supporting Scholarly Search by Query Expansion and Citation Analysis," *Engineering, Technology & Applied Science Research*, vol. 10, no. 4, pp. 6102–6108, Aug. 2020, https://doi.org/10.48084/etasr.3655.

[25] K. Albishre, Y. Li, Y. Xu, and W. Huang, "Query-based unsupervised learning for improving social media search," *World Wide Web*, vol. 23, no. 3, pp. 1791–1809, May 2020, https://doi.org/10.1007/s11280-019-00747-0.

[26] E. Voorhees, N. Craswell, B. Mitra, D. Campos, and E. Yilmaz, "Overview of the TREC 2019 Deep Learning Track," National Institute of Standards and Technology, SP1250, 2020.

[27] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos, "Overview of the TREC 2020 deep learning track." arXiv, 2021, https://doi.org/10.48550/ARXIV.2102.07662.

[28] P. Bajaj *et al.*, "MS MARCO: A Human Generated MAchine Reading COmprehension Dataset." arXiv, Oct. 31, 2018, https://doi.org/10.48550/arXiv.1611.09268.

[29] J. Lin, X. Ma, S. C. Lin, J. H. Yang, R. Pradeep, and R. Nogueira, "Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, July 2021, pp. 2356–2362, https://doi.org/10.1145/3404835.3463238.

[30] A. Q. Jiang *et al.*, "Mistral 7B." arXiv, Oct. 10, 2023, https://doi.org/10.48550/arXiv.2310.06825.

[31] R. Nogueira, Z. Jiang, R. Pradeep, and J. Lin, "Document Ranking with a Pretrained Sequence-to-Sequence Model," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Aug. 2020, pp. 708–718, https://doi.org/10.18653/v1/2020.findings-emnlp.63.

[32] S. Khalid, S. Wu, and F. Zhang, "A multi-objective approach to determining the usefulness of papers in academic search," *Data Technologies and Applications*, vol. 55, no. 5, pp. 734–748, Oct. 2021, https://doi.org/10.1108/DTA-05-2020-0104.