Query-Type Sensitivity and Selective Benefits of Hybrid Lexical–Semantic Retrieval: An Experimental Analysis with BM25 and Multilingual SBERT.

Abstract

Hybrid lexical–semantic retrieval combines signals derived from explicit term matching and dense semantic representations. However, an aggregate retrieval score can conceal heterogeneous behavior across information needs. This study examines that heterogeneity at query and predefined query-type levels using a controlled university information-retrieval benchmark of 50 queries and 19 documents, corresponding to 950 query–document pairs. Four retrieval conditions are evaluated descriptively: BM25, multilingual SBERT, Reciprocal Rank Fusion (RRF), and normalized lexical–semantic score fusion with a fixed weight $\alpha=0.5$. The primary metric is nDCG@5. The fixed-weight hybrid achieves a descriptive mean nDCG@5 of 0.969407, compared with 0.900470 for BM25, 0.823640 for SBERT, and 0.907719 for RRF. At the query level, the hybrid yields a higher nDCG@5 than BM25 for 10 queries, ties for 38, and is lower for 2. The largest descriptive subgroup gain occurs for PARAPHRASE queries ($\Delta=+0.2054$, $n=12$), whereas DISCRIMINATIVE queries provide a counterexample (BM25=0.9180; hybrid=0.9167). The evidence therefore supports a bounded conclusion: hybrid benefit is selective and varies across query types in the tested benchmark. The study does not establish statistical significance within subgroups, a universal optimal fusion weight, causal mechanisms, or external generalization.

Key words: -
hybrid retrieval; query-type sensitivity; selective benefit; BM25; multilingual SBERT; nDCG@5

…….....

Introduction: -

Information retrieval systems increasingly combine lexical and semantic signals to address information needs expressed with different degrees of lexical overlap. Lexical retrieval remains effective when relevant documents share explicit terms with the query, with BM25 providing a well-established and widely used ranking baseline [1]. Dense sentence representations provide a complementary mechanism by capturing semantic relatedness beyond exact lexical overlap [2], [3]. More recent neural retrieval approaches, including late-interaction models, have further demonstrated the potential of learned representations for matching queries and documents at different levels of semantic and contextual correspondence [4], [5]. Rank aggregation provides an alternative mechanism for combining retrieval systems without requiring their underlying score distributions to be directly calibrated [6].

However, the question of whether hybrid retrieval is beneficial cannot be adequately characterized by a single aggregate effectiveness measure. An improvement in mean retrieval performance may result from a combination of substantial gains for some queries, unchanged performance for others, and degradations for a smaller subset. Consequently, two systems with similar or different aggregate scores may exhibit substantially different query-level behavior. This issue is particularly relevant when the evaluated queries represent heterogeneous information needs or differ in their linguistic formulation. Benchmarking studies have shown that retrieval effectiveness can vary substantially across evaluation settings, query characteristics, and retrieval tasks [7].

The scientific problem addressed in this study is therefore the characterization of query-level and query-type heterogeneity in hybrid lexical–semantic retrieval. Rather than focusing exclusively on whether fusion produces a higher overall score, the study investigates whether the observed benefit is distributed uniformly across queries or concentrated within specific types of information needs. This perspective makes it possible to distinguish an aggregate improvement from a selective improvement and to identify query types for which lexical and semantic signals appear to provide complementary retrieval evidence.

The present study consequently uses query type as an analytical lens for examining the behavior of a lexical–semantic retrieval hybrid. The objective is not to introduce a new retrieval architecture or fusion algorithm, but to

characterize the behavior of a fixed α=0.5 lexical–semantic combination relative to established retrieval baselines. The analysis examines overall retrieval effectiveness, query-level differences relative to BM25, variation across the predefined query taxonomy, and sensitivity to the fusion-weight grid. Particular attention is given to identifying query types for which the hybrid exhibits the largest descriptive gains as well as cases in which fusion does not improve retrieval effectiveness.

The evaluation is conducted on a controlled experimental universe comprising 50 queries and 19 documents, resulting in 950 query–document pairs. The relevance judgments and experimental results used in the analysis are drawn from the audited benchmark and are treated as a fixed evaluation setting. The study therefore makes benchmark-specific descriptive claims rather than claims of statistical superiority, universal optimality of the α=0.5 configuration, or generalization to arbitrary information-retrieval environments.

The remainder of the paper is organized around four research questions.

Research questions

RQ1. How does retrieval effectiveness vary across the predefined query types?

RQ2. Which query types show the greatest descriptive benefit from lexical–semantic fusion?

RQ3. Is the observed hybrid benefit uniformly distributed across queries, or is it concentrated in a subset of queries?

RQ4. How does retrieval effectiveness vary with the fusion-weight parameter across the predefined query types?

Materials and Methods

Benchmark and ground truth

The controlled benchmark contains 50 queries and 19 documents, yielding 950 query–document pairs. The relevance ground truth contains 73 validated PRIMARY positive judgments. All four retrieval conditions are evaluated on the same query–document universe, so comparisons are not confounded by different candidate collections.

Query taxonomy

Table 1. Distribution of the predefined query types.

Query type	n	Proportion
COMPARATIVE	2	0.0400
DEFINITIONAL	1	0.0200
DIRECT	13	0.2600
DISCRIMINATIVE	6	0.1200
MULTI_CONSTRAINT	16	0.3200
PARAPHRASE	12	0.2400

Retrieval conditions

BM25 is used as the lexical baseline [1]. The semantic baseline uses sentence-transformers/paraphrase-multilingual-mpnet-base-v2. RRF is used as a rank-based comparator with k=60 [2]. The hybrid condition combines independently normalized BM25 and SBERT scores using a convex weight α.

Score normalization and fusion

For each query, the score produced by retrieval method m for document d is normalized by query-wise min–max scaling:

$$\tilde{S}_m(q,d) = \frac{[S_m(q,d) - Smin_{,m}(q)]}{[Smax_{,m}(q) - Smin_{,m}(q)]} \quad (1)$$

The hybrid score is then obtained by convex combination of the normalized lexical and semantic scores. The predefined grid is α∈ {0.0,0. 1…,1.0}; α=1 corresponds to BM25 and α=0 to SBERT.

$$H\alpha(q,d) = \alpha\, \tilde{S}BM25(q,d) + (1 - \alpha)\tilde{S}SBERT(q,d) \quad (2)$$

RRF combines the ranks produced by the component systems, with k fixed at 60:

$$RRF(d) = \Sigma r\ 1-[k + rankr(d)],\ k = 60 \quad (3)$$

Evaluation metrics

nDCG@5 is the primary metric because the analysis concerns the quality of the Top-5 ranking. Precision@5 and Recall@5 are retained as secondary descriptive measures. The same relevance judgments and cutoff are applied to every retrieval condition.

$$nDCG@k = DCG@k-IDCG@k \quad (4)$$

Query-level and query-type analysis

For each query, the difference Δ= nDCG@5(Hybrid α=0.5) − nDCG@5(BM25) is computed. A query is classified as improved, equal, or degraded according to the sign of Δ. Subgroup means are then calculated for each predefined QueryType. No inferential test is performed separately within query types; COMPARATIVE (n=2) and DEFINITIONAL (n=1) are explicitly treated as descriptive.

$$DCG@k = \Sigma_{\mathrm{i}}^{=1\,\mathrm{k}} \frac{(2^{rel}{}_{\mathrm{i}} - 1)}{log^2(i + 1)} \quad (5)$$

Resultats

Aggregate descriptive performance

Table 2. Mean nDCG@5 across the evaluated retrieval conditions.

Condition	Mean nDCG@5	Role
Hybrid α=0.5	0.969407	Primary hybrid condition
BM25	0.900470	Lexical baseline
RRF (k=60)	0.907719	Rank-fusion comparator
Multilingual SBERT	0.823640	Semantic comparator

Retrieval effectiveness by query type

Table 3. Mean nDCG@5 by predefined query type.

Query type	n	SBERT	BM25	Hybrid α=0.5	Δ Hybrid−BM25
COMPARATIVE	2	1.0000	1.0000	1.0000	+0.0000
DEFINITIONAL	1	0.6309	1.0000	1.0000	+0.0000
DIRECT	13	0.7254	0.9432	0.9716	+0.0284
DISCRIMINATIVE	6	0.8333	0.9180	0.9167	-0.0014
MULTI_CONSTRAINT	16	0.8863	0.9612	1.0000	+0.0388
PARAPHRASE	12	0.8282	0.7396	0.9450	+0.2054

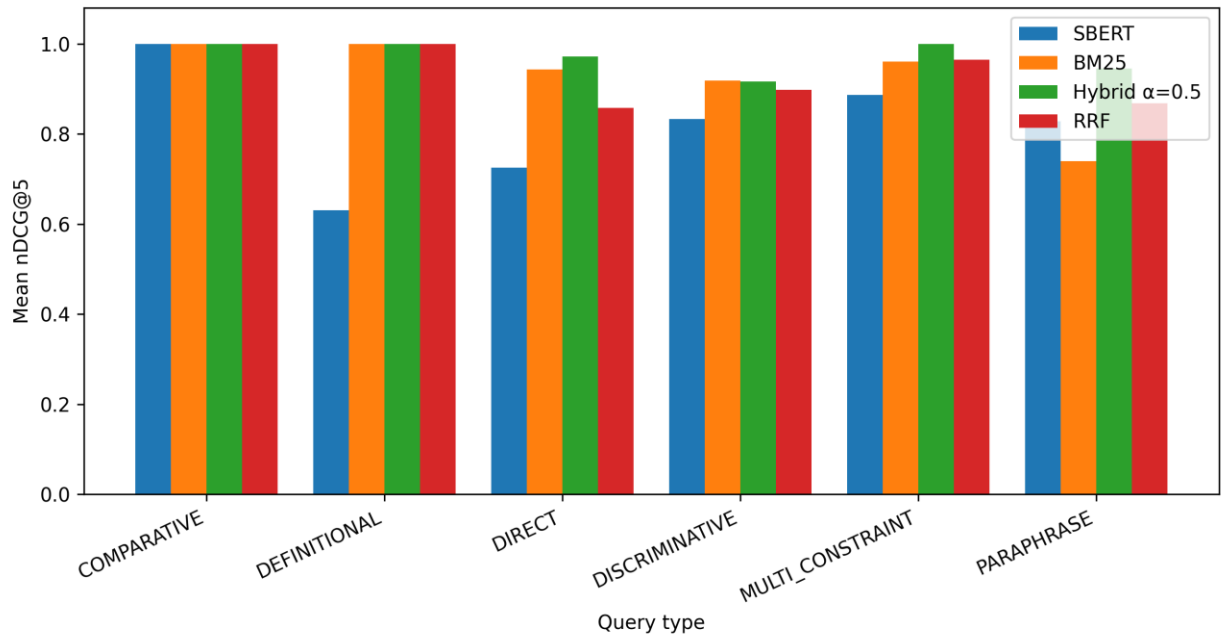

Figure 1. Descriptive nDCG@5 by query type and retrieval condition.

Query-level selective benefit

Table 4. Query-level outcomes of the fixed α=0.5 hybrid relative to the reference conditions.

Comparison	Improved	Equal	Degraded
Hybrid α=0.5 vs BM25	10	38	2
Hybrid α=0.5 vs RRF	10	40	0

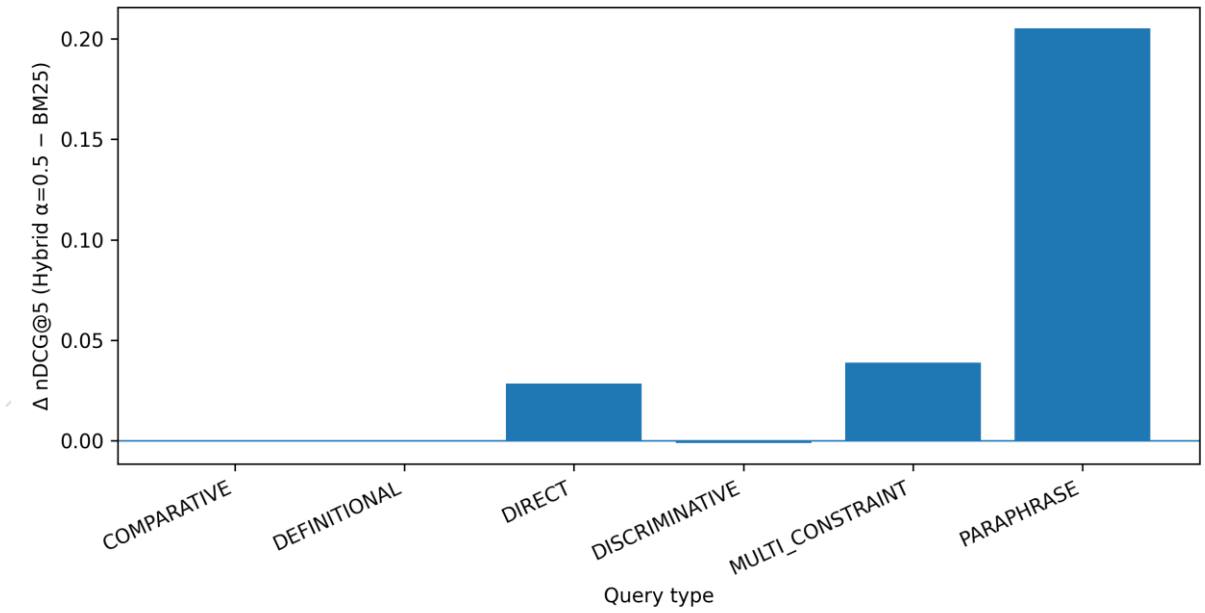

Figure 2. Query-type differences between the fixed α=0.5 hybrid and BM25.

The query-level comparison against BM25 shows a non-uniform pattern: the hybrid is higher for 10 queries, equal for 38, and lower for 2. This distribution is the operational basis for the term selective benefit used in this manuscript. It does not imply that the hybrid is generally inferior or superior; it indicates that most queries are tied at the reported cutoff while a smaller subset shows measurable differences.

Fusion-weight sensitivity by query type

The predefined α grid produces different response profiles across query types. DIRECT reaches 1.0 at α=0.6; MULTI_CONSTRAINT reaches 1.0 at α=0.5; PARAPHRASE reaches its maximum at α=0.5; DISCRIMINATIVE reaches its highest value at α=1.0 in the audited grid. COMPARATIVE and DEFINITIONAL are too small to support a meaningful sensitivity interpretation.

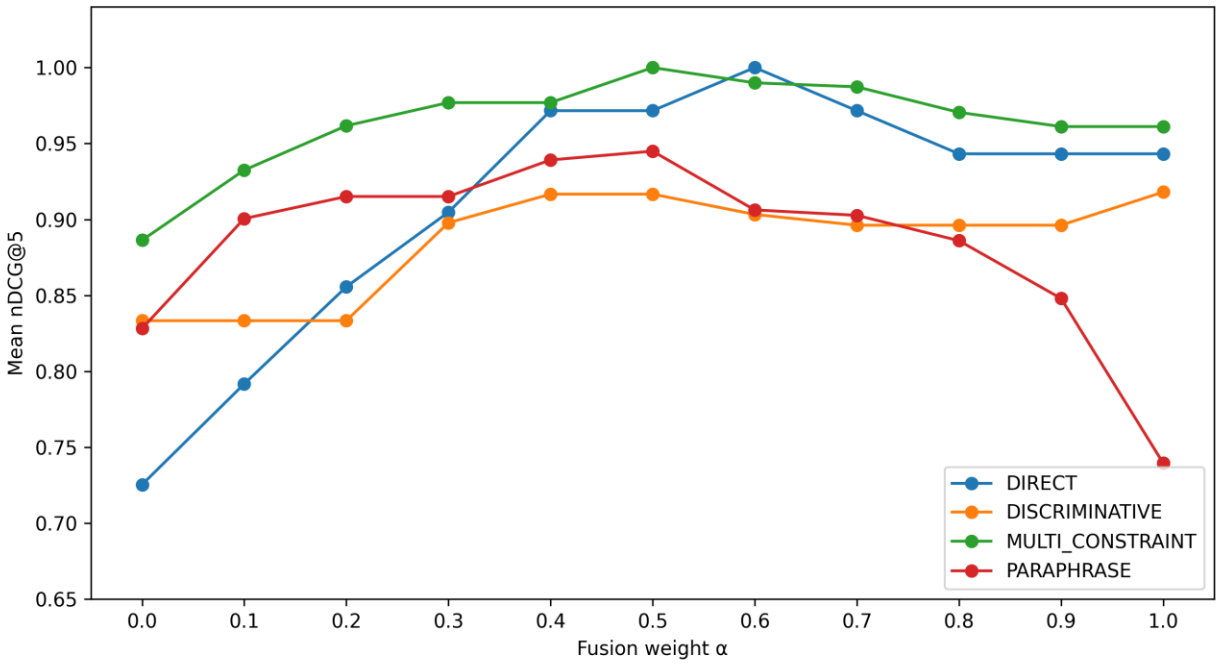

Figure 3. Fusion-weight sensitivity of mean nDCG@5 for the four query types with sufficient subgroup size.

Discussion

Selective rather than uniform benefit

The principal finding is not simply that the hybrid has a higher aggregate score. The query-level evidence shows that the observed benefit is concentrated in a subset of queries. The largest subgroup difference occurs for PARAPHRASE queries, where the hybrid mean exceeds BM25 by 0.2054 nDCG@5 points (n=12). This is consistent with the interpretation that lexical–semantic combination can be particularly useful when relevant information is expressed with alternative wording. This remains an interpretation of the observed pattern, not a causal finding.

Query-type heterogeneity

The subgroup results show that the direction and magnitude of the hybrid–BM25 difference are not identical across query types. DIRECT and MULTI_CONSTRAINT queries show positive descriptive differences, PARAPHRASE shows the largest positive difference, and DISCRIMINATIVE provides a counterexample in which BM25 is slightly higher. The COMPARATIVE and DEFINITIONAL groups are too small to support stable subgroup conclusions.

Fusion-weight sensitivity

The α curves indicate that the best descriptive weight depends on query type. This observation argues against interpreting α=0.5 as a universal optimum. In this study, α=0.5 is a fixed analytical condition used for the primary query-level comparison, while the grid analysis is used to characterize sensitivity.

Implications for retrieval evaluation

The results indicate that reporting only an aggregate retrieval metric can conceal important query-level structure. A query-type-aware evaluation can reveal where a hybrid method adds measurable value and where a lexical baseline already performs strongly. These implications concern evaluation and diagnosis; they do not justify dynamic query routing or production deployment without external validation.

Threats to validity and limitations

The evaluation is conducted on a controlled benchmark of 50 queries and 19 documents, which limits external generalization to larger, more diverse retrieval collections and information needs.

Query-type groups are unbalanced; COMPARATIVE (n=2) and DEFINITIONAL (n=1) are too small to support reliable subgroup inference. Their results are therefore interpreted descriptively only.

Only one multilingual SBERT encoder configuration is evaluated; therefore, the findings cannot be assumed to represent the behavior of other dense encoders or embedding models.

The hybrid relies on query-wise min–max normalization and a predefined α grid; alternative normalization, calibration, or fusion-weight choices may produce different retrieval profiles.

RRF is evaluated with k=60; therefore, the comparison characterizes this predefined RRF configuration rather than the full range of possible RRF parameterizations.

The reported results are descriptive. No inferential statistical testing is performed within query-type subgroups, and no claim of universal superiority or universal optimality is made.

Conclusion

This study examined hybrid lexical–semantic retrieval through query-level and query-type analysis on a controlled benchmark. The fixed α=0.5 hybrid achieved the highest descriptive mean nDCG@5 among the four evaluated retrieval conditions. However, the query-level comparison shows that the observed benefit is selective: 10 queries improved over BM25, 38 were tied, and 2 were lower. PARAPHRASE queries exhibited the largest descriptive subgroup gain, whereas DISCRIMINATIVE queries provided a counterexample with a small negative difference. These findings show that aggregate retrieval effectiveness can conceal heterogeneous query-level behavior and support the use of query-level and query-type analyses when interpreting hybrid retrieval performance. The conclusions remain bounded by the evaluated benchmark, retrieval configurations, relevance judgments, and predefined fusion-weight grid.

References

1. S. Robertson and H. Zaragoza, “The Probabilistic Relevance Framework: BM25 and Beyond,” Foundations and Trends in Information Retrieval, vol. 4, no. 1–2, pp. 1–174, 2009, doi: 10.1561/1500000019.

2. G. V. Cormack, C. L. A. Clarke, and S. Büttcher, “Reciprocal Rank Fusion outperforms Condorcet and Individual Rank Learning Methods,” in Proc. SIGIR, 2009, pp. 758–759, doi: 10.1145/1571941.1572114.

3. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” in Proc. EMNLP-IJCNLP, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.

4. V. Karpukhin et al., “Dense Passage Retrieval for Open-Domain Question Answering,” in Proc. EMNLP, 2020, pp. 6769–6781, doi: 10.18653/v1/2020.emnlp-main.550.

5. O. Khattab and M. Zaharia, “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,” in Proc. SIGIR, 2020, pp. 39–48, doi: 10.1145/3397271.3401075.

6. N. Thakur et al., “BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models,” in Proc. NeurIPS Datasets and Benchmarks Track, 2021.

7. T. Formal, B. Piwowarski, and S. Clinchant, “SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking,” in Proc. SIGIR, 2021, pp. 2288–2292, doi: 10.1145/3404835.3463098.

8. O. Ram et al., “Learning to Retrieve Passages without Supervision,” in Proc. NAACL-HLT, 2022, pp. 2687–2700, doi: 10.18653/v1/2022.naacl-main.193.

9. C. Xu et al., “LaPraDoR: Unsupervised Pretrained Dense Retriever for Zero-Shot Text Retrieval,” Findings of ACL, 2022, pp. 3557–3569, doi: 10.18653/v1/2022.findings-acl.281.

10. W. X. Zhao et al., “Dense Text Retrieval Based on Pretrained Language Models: A Survey,” ACM Transactions on Information Systems, vol. 42, no. 4, Art. 89, 2024, doi: 10.1145/3637870.

Declarations

Author contributions: Mamadou B: Conceptualization, Methodology, Investigation, Writing — original draft. Tiemoman K: Conceptualization, Supervision, Writing — review & editing. Yao K: Software, Formal analysis, Data curation, Visualization, Writing — review & editing.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of interest: The authors declare no known competing financial interests or personal rela-tionships that could have appeared to influence the work reported in this paper.

Data availability: The data that support the findings of this study are available from the corresponding author upon reasonable request.