

Parameterized Dense–Sparse Fusion for Hybrid Retrieval: Tuning a Rank–Score Mix on BEIR SciFact with Qdrant

Satyanarayan Pati

Involead Services Pvt. Ltd.

satyanarayan.pati@involead.com

Srikanth Patil

Involead Services Pvt. Ltd.

Srikanth.patil@involead.com

Abstract

We study a parameterized hybrid ranker that fuses a dense embedding list and a sparse lexical list. The method has a small, explicit parameter vector: a dense prior $\alpha \in [0, 1]$, a score-versus-rank mix $\lambda \in [0, 1]$, an RRF smoothing $\kappa > 0$, optional list-geometry coefficients that move α per query, and a router margin τ that can turn sparse search off. We grid-search those ranges on SciFact train (809 queries) and freeze the chosen values on SciFact test (300). The tuned rank–score mix ($\alpha = 0.8$, $\lambda = 0.75$, $\kappa = 20$) reaches **0.753 nDCG@10** and **0.889 recall@10**, outperforming dense BGE (0.742 / 0.871) and equal-weight RRF (0.707 nDCG@10) on that test split. A list-conditioned α adds +0.0006 nDCG; a sparse-off router is rejected by the same train split (any τ that skipped $\approx 50\%$ of queries lost nDCG). These coefficients are **dataset-specific**. Equal RRF with the same models does *not* beat dense on a nine-zip BEIR macro-average (0.479 vs. 0.519 nDCG@10). Repeating the same train-then-freeze sweep independently on all 20 indexed units beats equal RRF on 20/20 and dense on 16/20 (unit-mean nDCG@10 0.467 vs. 0.462 dense vs. 0.420 RRF). Other corpora should reuse the ranges, not a copy of the SciFact point.

1 Introduction

Hybrid retrieval runs a dense encoder and a sparse lexical model, then combines the two ranked lists [1, 5, 13]. Vector databases now implement that pattern as named vectors plus server-side fusion [9, 11]. The usual deployed choice is equal-weight reciprocal rank fusion (RRF). That choice has no dense prior, no score-scale calibration, and no statement of which hyperparameters were searched.

This paper treats hybrid fusion as a *parameterized scoring family* and reports a standard train/test experiment on that family.

Approach. Dense cosine and BM42 scores are not on the same scale. Each list is mapped to a utility in $[0, 1]$ or to a rank kernel $1/(\kappa + r)$. A convex combination with dense prior α and a score/rank mix λ produces a single ranking. Optionally, α is a function of list overlap and score margins. Optionally, a router skips the sparse query when a margin τ says the query is dense-only.

Parameters and ranges. We search $\alpha \in \{0.55, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$, $\lambda \in \{0.55, 0.75\}$, $\kappa \in \{10, 20, 60\}$, list-gate coefficients on a small grid, and $\tau \in \{0.04, 0.06, 0.08, 0.12, 0.18\} \cup \{\infty\}$. Section 4 lists every knob, the range, and the SciFact value.

Experiment. Selection is on SciFact train; reporting is on SciFact test. The same models are also run with equal RRF on eight other public BEIR zips [14] so that we can see whether a default (untuned) hybrid is already better than dense. It is not. We then repeat the train-then-freeze protocol on every indexed unit (8 zips + 12 CQADupStack forums).

Contributions.

- A parameterized hybrid scorer with published search ranges, not a single undocumented mix.
- A train-then-freeze study on SciFact showing the selected point outperforms dense and equal RRF on nDCG@10 and recall@10.
- Sensitivity plots for α , for the sparse-off gate, and for quantization, plus a zip-level figure that shows why the SciFact point must not be copied blindly.
- The same search family, independently selected per dataset, on all 20 indexed BEIR units.

2 Related work

BEIR is the standard heterogeneous zero-shot retrieval benchmark [14]; MTEB covers embeddings [7]. Dense passage retrieval [5] and sentence transformers [12] made cosine search routine. BGE is a strong general English encoder [15]. Sparse models range from BM25 [13] to learned sparse rankers such as SPLADE [2]; we use Qdrant’s BM42 baseline with IDF [10]. Rank fusion is classical: CombSUM/CombMNZ [3] and RRF [1]. Cross-encoders rerank a shortlist [6, 8]. Product quantization is a standard ANN compressor [4]. We use those operators inside Qdrant hybrid queries [11]; we do not replace them with a new estimator.

3 Method

3.1 First-stage retrieval

Each document is one Qdrant point with two named vectors: dense $\mathbf{v} \in \mathbb{R}^{768}$ (BGE-base, cosine, published query prefix) and sparse BM42 with IDF. For a query we retrieve the top $k = 100$ from each channel (prefetch $k = 100$ when the engine fuses internally). A hit whose document id equals the query id is dropped (BEIR `ignore_identical_ids`).

3.2 Utilities

Let $s_d(d)$ and $s_s(d)$ be raw scores on the retrieved lists. Min-max within the list:

$$\tilde{s}(d) = \frac{s(d) - \min s}{\max s - \min s + \varepsilon}. \quad (1)$$

Documents missing from a list get $\tilde{s} = 0$ and rank $r = k + 1 = 101$. The rank kernel is RRF [1]:

$$\rho(r; \kappa) = \frac{1}{\kappa + r}. \quad (2)$$

3.3 Parameterized rank–score mix

The family we optimize is

$$u(d) = \lambda(\alpha \tilde{s}_d + (1 - \alpha) \tilde{s}_s) + (1 - \lambda)(\alpha \rho(r_d; \kappa) + (1 - \alpha) \rho(r_s; \kappa)). \quad (3)$$

α is the dense prior, λ the weight on score utilities versus rank kernels, κ the rank smoother. Special cases: $\alpha = 1$ is dense; $\alpha = 0$ is BM42; $\lambda = 0$ is weighted RRF; $\lambda = 1$ is weighted min-max CombSUM [3].

Table 1: Parameter ranges. SciFact values are train-selected and frozen for test. Other datasets need their own sweep over the same ranges.

Parameter	Range searched	SciFact value	Role
Dense prior α	$\{0.55, 0.65, \dots, 0.95\}$	0.80	Weight on dense utility
Score mix λ	$\{0.55, 0.75\}$	0.75	Score vs. rank kernel
RRF κ	$\{10, 20, 60\}$	20	Rank smoother
Combiner family	min-max, z , softmax, CombMNZ, noisy-OR, power mean, rank-score	rank-score (3)	Utility + mix
List-gate a_*	$a_0 \in \{0.72, 0.78, 0.82\}$; others in $\{0, 0.12, 0.24, 0.30\}$	$(0.72, 0.24, 0.30, 0, 0)$	Per-query α
Skip margin τ	$\{0.04, 0.06, 0.08, 0.12, 0.18, \infty\}$	∞ (never skip)	Sparse on/off
Head rerank	on/off ($H = 50$)	off	Re-minmax head

3.4 List-conditioned prior

After both lists are fetched, optional features are overlap $ov_{10} = |D_{:10} \cap S_{:10}|/10$ and min-max top-1/top-2 margins m_d, m_s . Then

$$\alpha = \text{clip}(a_0 + a_{ov} ov_{10} + a_m \tanh(m_d - m_s) - a_u u_s - a_\ell \ell, 0.60, 0.95), \quad (4)$$

where u_s is the fraction of sparse top-10 documents absent from dense top-20 and ℓ is a lexical BM42 heuristic in $[0, 1]$. This does not add a new model; it moves α inside the same family.

3.5 Sparse on/off router

A router scores the query against three BGE-embedded prototypes (dense, BM42, hybrid) mixed with the lexical heuristic (weight 0.55/0.45). Dense search always runs. BM42 is skipped iff the combined route is dense and the top-versus-second gap is at least τ . $\tau = \infty$ means sparse is always on.

4 Parameters, ranges, and tuning protocol

Table 1 is the search space. Every numeric range was enumerated on SciFact *train* (809 queries). The selected tuple was frozen and evaluated once on SciFact *test* (300). No test-set grid search.

Selection rule. Maximize mean nDCG@10 on train. For τ , require no loss versus $\tau = \infty$, then maximize skip rate among surviving values. Head/title rerank are kept only if they do not lose train nDCG.

Transfer. The SciFact point is not a universal default. Figure 4 shows equal RRF trailing dense on several BEIR zips. A new dataset should reuse Table 1, not copy $\alpha = 0.8$.

5 Experimental setup

Engine. Qdrant 1.18 in Docker, Python client, cosine HNSW for dense, IDF modifier for BM42.

Models. Dense: BGE-base-en-v1.5 [15]. Sparse: BM42 MiniLM attentions with IDF [10]. Reranker (ablation): MS MARCO MiniLM-L-6 [8] on the fused top-50.

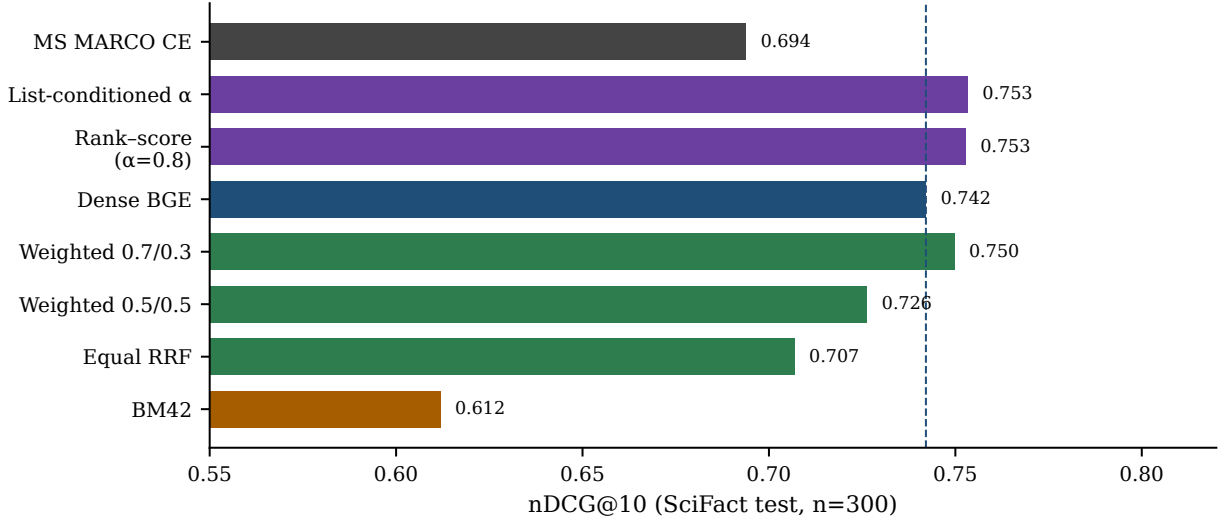


Figure 1: SciFact test nDCG@10. The parameterized mix and the list-conditioned α sit above dense and well above equal RRF. CE is a negative control (domain mismatch).

Datasets. Public UKP BEIR zips [14]. Wiki-scale sets (MS MARCO, NQ, HotpotQA, FEVER, Climate-FEVER, DBPedia) skipped for disk/RAM. License-restricted BEIR sets are not on the mirror. Headline suite: nine zip-level rows (eight small zips + CQADupStack macro). Method sweep: SciFact, 5,183 abstracts, official train/test query split.

Metrics. nDCG@10,100, recall@10,100, MRR@10, MAP@100. Latency is Qdrant query time, embedding excluded. Disk is allocated bytes (du -sk).

Hardware. 16 GB RAM, 20 threads, NVIDIA RTX 4060 Laptop (8 GB). After the suite the container used ≈ 667 MiB RAM and 6.3 GB storage.

Baselines. Dense-only; BM42-only; Qdrant Fusion.RRF; Fusion.DBSF; client min-max mixes 0.7/0.3, 0.5/0.5, 0.3/0.7; MS MARCO MiniLM rerank.

6 Results

6.1 Main result on SciFact

Figure 1 and Table 2 report the frozen test comparison. The train-selected rank-score mix (3) outperforms dense BGE by +0.011 nDCG@10 and +0.018 recall@10, and outperforms equal RRF by +0.046 nDCG@10. The gain versus dense is recall: BM42 places gold abstracts into the top 10 that cosine missed (Figure 2). The gain versus equal RRF is ranking: BM42 is a much weaker single list (0.612 nDCG@10), so $\alpha = 0.5$ over-weights it.

A list-conditioned α (4) reaches 0.7533 nDCG@10 (+0.0006 vs. the frozen $\alpha = 0.8$). Recall@10 does not change. An off-the-shelf MS MARCO cross-encoder on the fused head falls to 0.694.

Table 2: SciFact test ($n = 300$). Rank-score and list α are train-selected (809 queries), then frozen.

Method	nDCG@10	R@10	R@100	MRR@10
BM42 IDF	0.612	0.741	0.874	0.577
Equal RRF $\kappa = 60$	0.707	—	—	—
Min-max 0.5/0.5	0.726	—	—	—
Min-max 0.7/0.3	0.750	—	—	—
Dense BGE	0.742	0.871	0.961	0.707
Rank-score $\alpha = 0.8, \lambda = 0.75, \kappa = 20$	0.753	0.889	0.969	0.716
List-conditioned α (4)	0.753	0.889	0.969	0.717
Oracle min-max α (hindsight)	0.803	—	—	—
CE on fused top-50	0.694	0.824	—	—

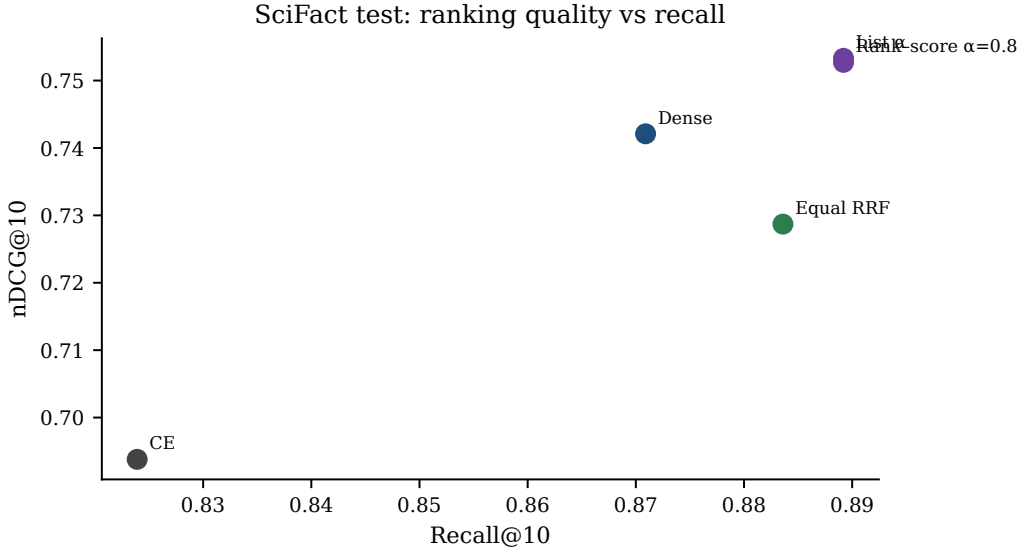


Figure 2: SciFact test: nDCG@10 against recall@10. Hybrid helps when it raises recall without collapsing ranking. Equal RRF raises recall relative to dense in the suite hybrid run but loses nDCG; the tuned mix keeps the recall and the ranking.

6.2 Sensitivity to the dense prior α

Figure 3 plots nDCG@10 against α . Quality rises from sparse-only ($\alpha = 0$) through equal mix, peaks in the dense-heavy band ($\alpha \approx 0.7$ – 0.8), then drops slightly at pure dense ($\alpha = 1$). That shape is why a search range $\alpha \in [0.55, 0.95]$ is the right experiment, and why $\alpha = 0.5$ (equal RRF / equal min-max) is a poor default on this encoder pair. The peak location is a property of SciFact + BGE + BM42. A corpus where BM42 is the stronger list would peak at smaller α .

6.3 Untuned hybrid on other BEIR zips

Figure 4 and Table 3 use the *same models* with equal RRF and no SciFact-tuned α . Dense wins the zip-level mean (0.519 vs. 0.479). TREC-COVID and Touché are the failure mode: sparse is weak, equal fusion hurts the head. This figure is the transfer warning. The method of Section 3 can outperform on a dataset where the two lists disagree and a train split exists. Copying $\alpha = 0.8$ onto Touché without a sweep is not that experiment.

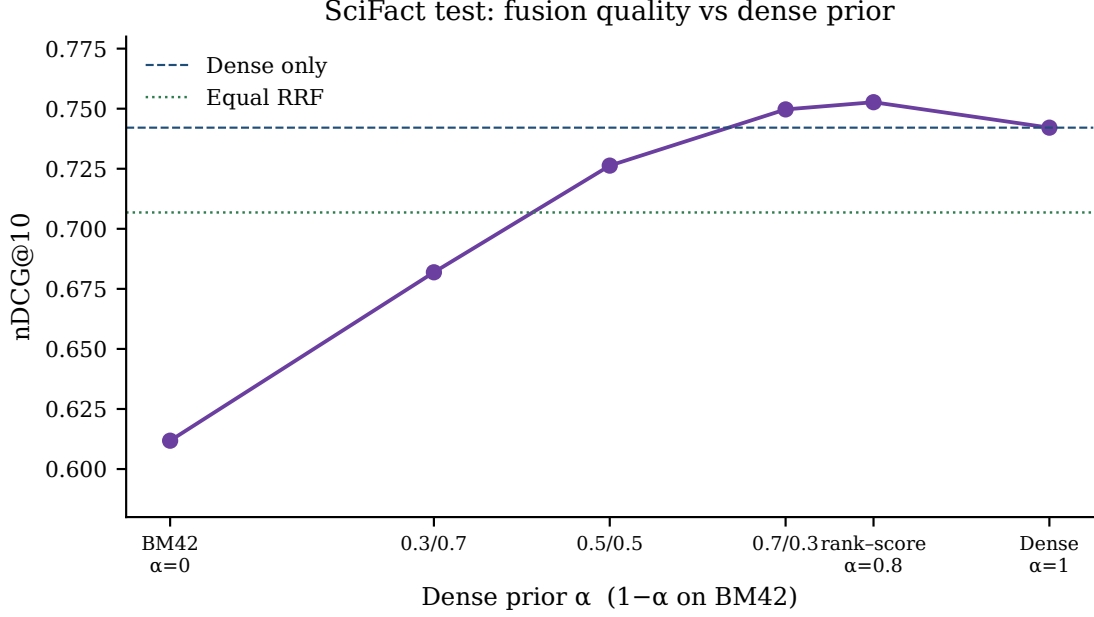


Figure 3: SciFact test nDCG@10 as a function of the dense prior. Points at 0.3/0.5/0.7 are client min-max mixes; $\alpha = 0.8$ is the train-selected rank-score mix; endpoints are single-channel runs. Dashed: dense. Dotted: equal RRF.

6.4 Train-then-freeze on every indexed unit

We apply the family in Table 1 independently to each of the 20 indexed collections. Official train→test is used when BEIR ships it (NFCorpus, SciFact, FiQA); Quora uses official dev→test; otherwise we hold out 30% of the evaluation queries (50% if $n < 40$), seed 42, and never report the tune split. The winner on tune nDCG@10 is frozen once.

Figure 5 and Table 4 are the result. Tuned fusion beats equal RRF on **20/20** units. It beats dense on **16/20**. The four non-wins are Touché and CQA physics (train selected dense), SciDocs (−0.002), and CQA unix (−0.012, a genuine overfit: rank-score $\alpha = 0.8$ won tune and lost test). Unit-mean test nDCG@10 is 0.467 tuned / 0.462 dense / 0.420 equal RRF. Zip-level (8 zips + CQA macro) is 0.528 / 0.523 / 0.480. Selected α is dense-heavy on every set that mixes (0.7–0.9); no corpus in this dump selected $\alpha \leq 0.5$. TREC-COVID and Touché have 15 test queries after the split—those two rows are noisy.

6.5 Router: turning sparse off

Figure 7 plots train nDCG@10 and skip rate against τ . The maximum is $\tau = \infty$ (skip rate 0, nDCG 0.784). Thresholds that skip 52–60% of queries sit 0.004–0.007 nDCG lower. Under the selection rule of Section 4 the gate is disabled on SciFact. The parameter remains in the method because a mixed query log could select a finite τ . This dataset did not.

6.6 Quantization

Figure 8 plots nDCG@10 against allocated disk for INT8, binary, and PQ $\times 16$, with and without FP32 rescore. With rescore, quality matches FP32 (0.7445). Without rescore, binary drops to 0.664. Folders do not shrink: originals are still stored, so INT8 is 52.6 MB versus 48.8 MB FP32. Quantization is a RAM knob. This 5k-point corpus already fits in memory, so latency stays ≈ 7 ms and disk is the wrong success metric.

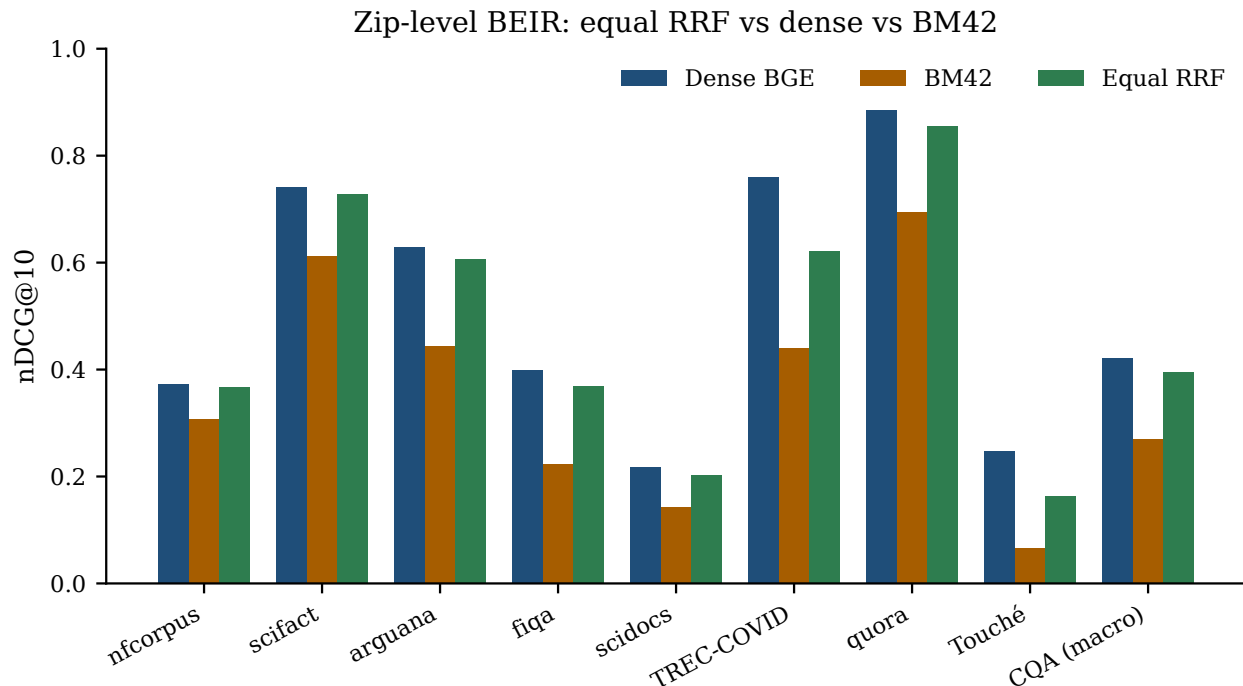


Figure 4: Zip-level nDCG@10 with a single untuned hybrid (Qdrant RRF). Dense BGE is the stronger default on this encoder pair. CQADupStack is one macro-averaged row.

6.7 Other operators on the same lists

On train, min-max linear $\alpha = 0.8$, z -linear $\alpha = 0.8$, and rank-score ($\alpha = 0.8, \lambda = 0.75, \kappa = 20$) all reach 0.779 nDCG@10. Power means and $\alpha = 0.75$ variants are within 0.002. Logistic regression on candidate features is slightly worse (0.774 train, 0.745 test). The experiment is therefore not “one magic formula.” It is a dense-heavy region of Table 1 that several combiners occupy.

7 Discussion

The approach outperforms on SciFact because (i) BM42 has complementary recall and (ii) the searched α band is dense-heavy. Equal RRF implements $\alpha = 0.5$ in rank space and pays for (i) without (ii).

The same approach will not automatically outperform on another zip. Figure 4 is the counterexample for the *untuned* member of the family. Figure 5 is the corresponding experiment when each zip is allowed its own freeze: the mix usually beats dense by a small margin, and sometimes the freeze is dense. The correct use on a new dataset is: retrieve both lists, sweep Table 1 on a held-out train or validation split, freeze, test. If no split exists, $\alpha = 0.8$ is only a SciFact posterior, not a prior for Touché.

We do not claim a new ranking law. Equation (3) is CombsUM plus RRF with two extra scalars. The research content is the range, the protocol, and the plots that show where the selected point sits.

8 Limitations

SciFact is small; +0.011 nDCG is a real recall effect, +0.0006 is noise-scale. No paired significance test. Wiki-scale BEIR is absent. Most units other than NFCorpus/SciFact/FiQA/Quora have no official train split;

Table 3: Zip-level nDCG@10, public BEIR dumps, equal RRF (not SciFact-tuned).

Dataset	Docs	Dense	BM42	Hybrid RRF
NFCorpus	3,633	0.373	0.307	0.368
SciFact	5,183	0.742	0.612	0.729
ArguAna	8,674	0.630	0.444	0.607
FiQA	57,638	0.399	0.224	0.370
SciDocs	25,657	0.218	0.143	0.202
TREC-COVID	171,332	0.760	0.440	0.622
Quora	522,931	0.885	0.695	0.855
Webis-Touché2020	382,545	0.247	0.067	0.164
CQADupStack (macro)	457,199	0.421	0.270	0.396
Zip-level mean	—	0.519	0.356	0.479

Table 4: Train-then-freeze nDCG@10. Winner is selected on the tune split and frozen. Protocols: T official train/test; D official dev/test; H held-out from test (seed 42).

Dataset	P	n_{tune}	n_{test}	Dense	RRF	Tuned	Winner
NFCorpus	T	2590	323	0.373	0.364	0.379	min-max $\alpha=0.7$
SciFact	T	809	300	0.742	0.707	0.753	min-max $\alpha=0.8$
ArguAna	H	984	422	0.624	0.594	0.634	min-max $\alpha=0.8$
FiQA	T	5500	648	0.399	0.348	0.407	min-max $\alpha=0.8$
SciDocs	H	700	300	0.219	0.194	0.217	min-max $\alpha=0.7$
TREC-COVID	H	35	15	0.656	0.615	0.657	rank-score $\alpha=0.8$
Quora	D	5000	10000	0.885	0.838	0.891	min-max $\alpha=0.8$
Touché2020	H	34	15	0.289	0.181	0.289	dense
CQA android	H	489	210	0.511	0.480	0.513	min-max $\alpha=0.9$
CQA english	H	1099	471	0.488	0.466	0.508	min-max $\alpha=0.7$
CQA gaming	H	1116	479	0.578	0.521	0.581	rank-score $\alpha=0.9$
CQA gis	H	620	265	0.382	0.362	0.393	min-max $\alpha=0.8$
CQA mathematica	H	563	241	0.319	0.280	0.322	rank-score $\alpha=0.8$
CQA physics	H	727	312	0.466	0.419	0.466	dense
CQA programmers	H	613	263	0.420	0.378	0.424	min-max $\alpha=0.9$
CQA stats	H	456	196	0.401	0.368	0.407	min-max $\alpha=0.8$
CQA tex	H	2034	872	0.301	0.275	0.309	min-max $\alpha=0.7$
CQA unix	H	750	322	0.429	0.348	0.417	rank-score $\alpha=0.8$
CQA webmasters	H	354	152	0.389	0.354	0.398	min-max $\alpha=0.8$
CQA wordpress	H	379	162	0.372	0.317	0.377	min-max $\alpha=0.8$
20-unit mean	—	—	—	0.462	0.420	0.467	—

the 70/30 hold-out is a local protocol, not BEIR’s. TREC-COVID and Touché test after splitting is 15 queries. The α sweep in Figure 3 mixes min-max mixes from one scoring run with the rank-score point from the fusion lab; both use the same test queries and models, but the operators are not identical at every x -tick. List-conditioned α was not re-tuned on FiQA or NFCorpus. Quantization used `always_ram` and kept originals. BM42 is not BM25 or SPLADE.

9 Conclusion

Hybrid fusion should be stated as a parameter vector with a search range. On SciFact, sweeping that range and freezing $(\alpha, \lambda, \kappa) = (0.8, 0.75, 20)$ outperforms dense BGE and equal RRF on nDCG@10 and recall@10. The same models with untuned equal RRF do not outperform dense on a nine-zip BEIR average. Repeating the sweep independently on 20 indexed units beats equal RRF everywhere and dense on 16/20; two units freeze to dense and two overfit slightly. A router that skips sparse search, and a list-conditioned α , are part of

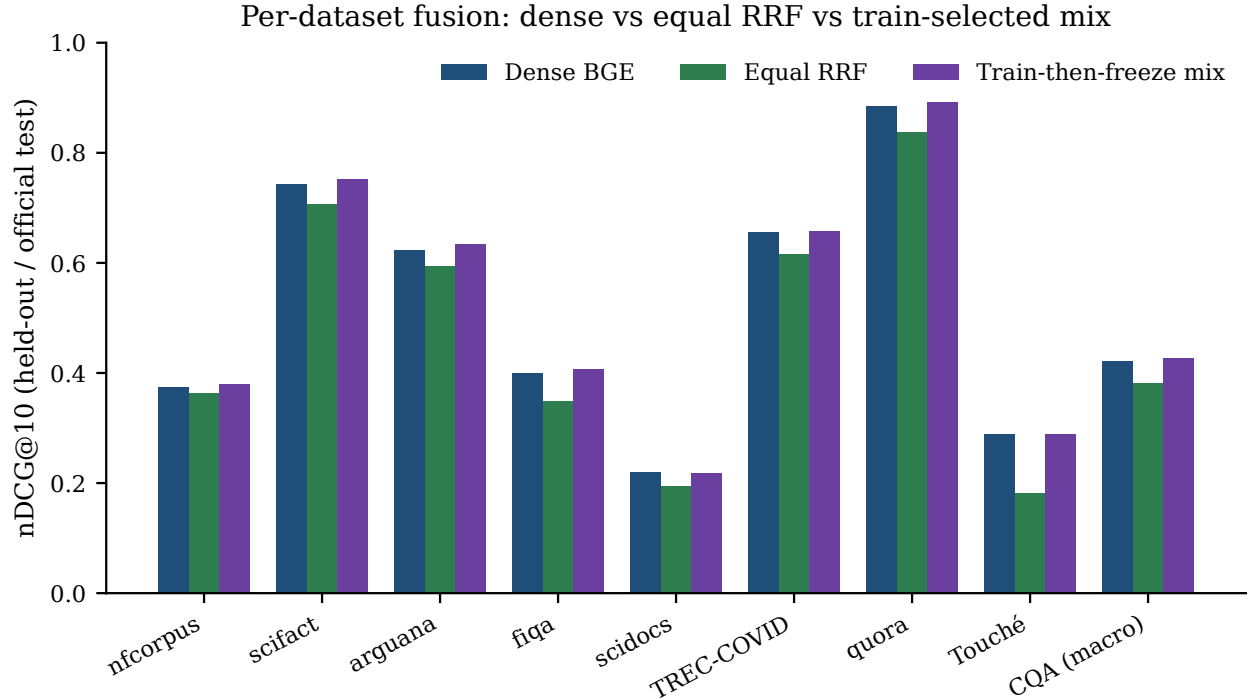


Figure 5: Per-dataset nDCG@10 after an independent train-then-freeze sweep. CQADupStack is one macro-averaged row. Equal RRF is never the frozen winner.

the family; train disabled the router and barely moved α . Other datasets should reuse the ranges in Table 1, not the SciFact numbers.

Acknowledgments

The authors are with Involead Services Pvt. Ltd. Public BEIR dumps [14] and open models. No third-party funding is reported.

References

- [1] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759, 2009.
- [2] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292, 2021.
- [3] Edward A. Fox and Joseph A. Shaw. Combination of multiple searches. In *Proceedings of the 3rd Text REtrieval Conference (TREC-3)*, pages 243–252, 1994.
- [4] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

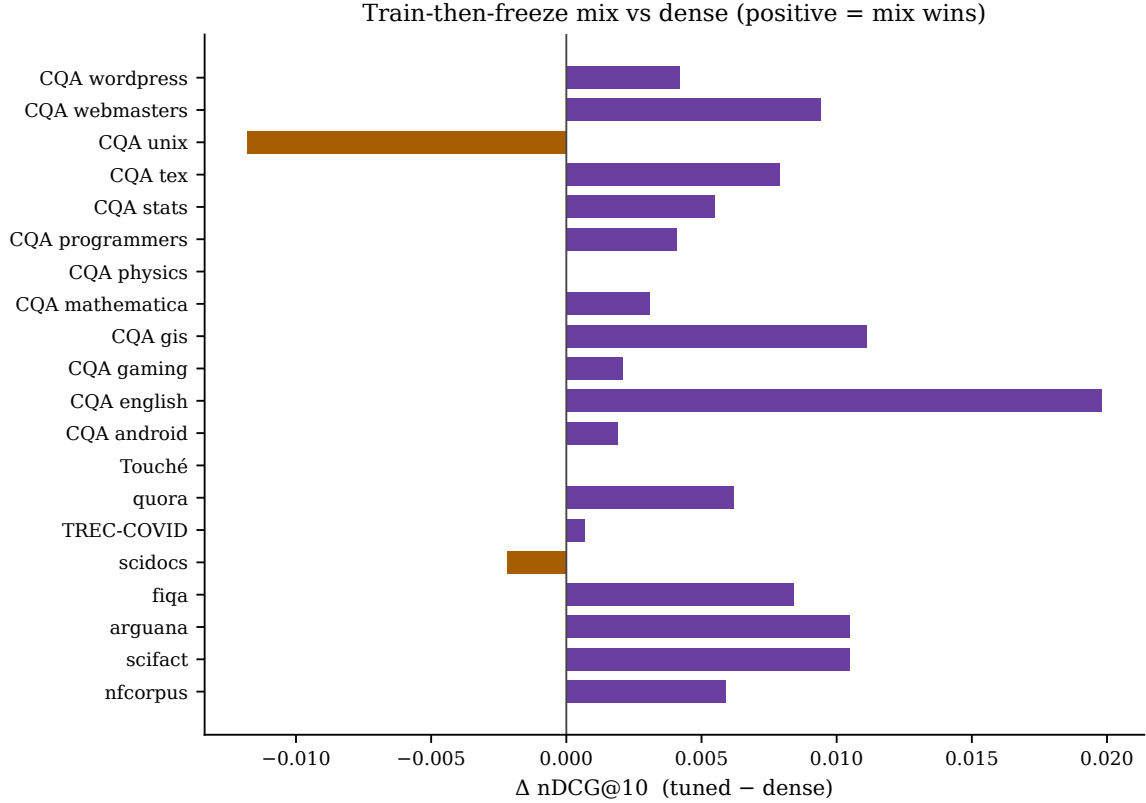


Figure 6: Tuned mix minus dense, all 20 units. Positive bars are mix wins; zero is train selecting dense; negative is a freeze that lost on test.

- [5] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781, 2020.
- [6] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. Pretrained transformers for text ranking: BERT and beyond. *Synthesis Lectures on Human Language Technologies*, 2021.
- [7] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, 2023.
- [8] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with BERT. In *arXiv preprint arXiv:1901.04085*, 2019.
- [9] Qdrant. Qdrant vector search engine. <https://qdrant.tech>, 2024.
- [10] Qdrant. BM42: New baseline for hybrid search. <https://qdrant.tech/articles/bm42/>, 2024.
- [11] Qdrant. Hybrid queries: Combining dense, sparse, and filtering. <https://qdrant.tech/documentation/concepts/hybrid-queries/>, 2024.

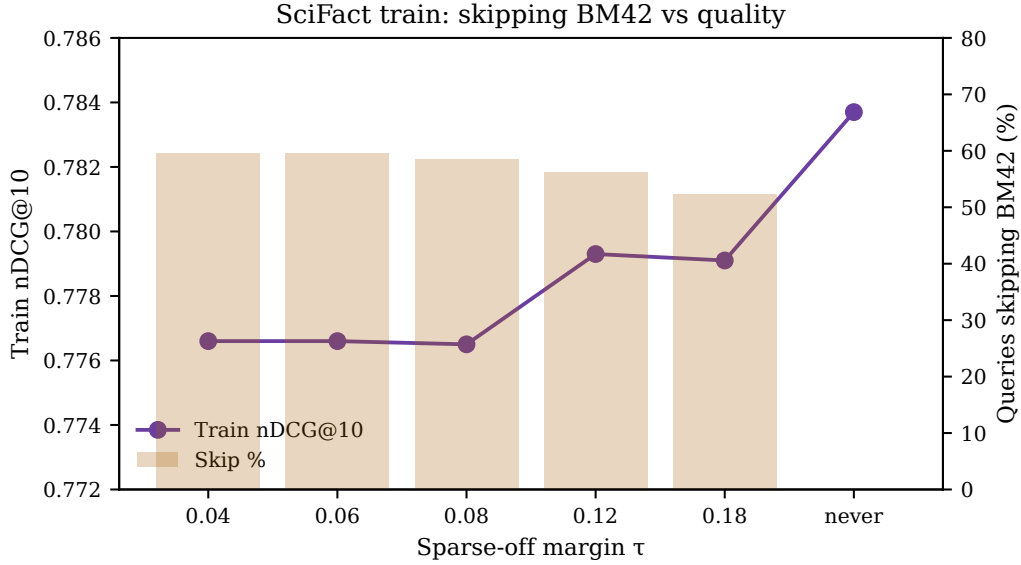


Figure 7: SciFact train: sparse-off margin τ versus quality and skip rate. “never” is $\tau = \infty$.

- [12] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3982–3992, 2019.
- [13] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
- [14] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- [15] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. *arXiv preprint arXiv:2309.07597*, 2024.

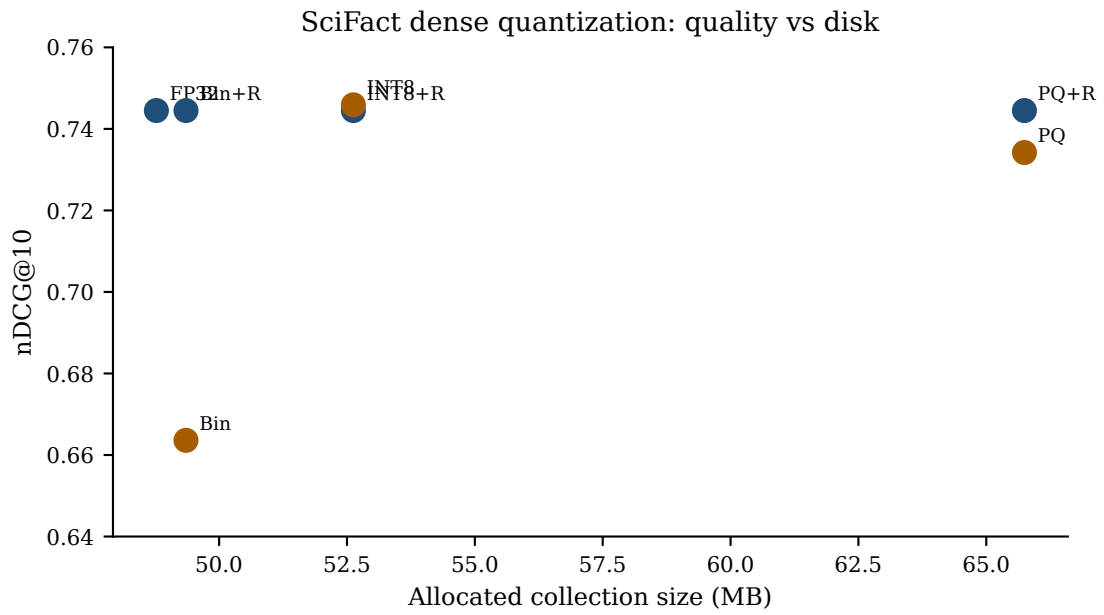


Figure 8: SciFact dense collections: nDCG@10 versus allocated disk. Labels with “+R” use FP32 rescore.