What Carries the Gain in LLM Query Rewriting for Dense Retrieval? A Mechanism Decomposition

Anonymous ACL submission

Abstract

We present a mechanism decomposition of LLM-based query rewriting for dense retrieval that isolates the contribution of each pipeline component. Reported gains on BEIR vary in sign across corpora and prior work has not attributed this variation to specific pipeline components. We ablate the injected term bank, the rewriter and its training, and the retrieval integration form on four BEIR corpora plus a held-out BioASQ probe, using BGE-M3 + BM25 RRF retrieval and paired-bootstrap CIs with multiplicity correction. The principal effect is carried by term content, not the rewriter or its training. Replacing extracted terms with count-matched random words degrades retrieval sharply, significantly so on NFCorpus and TC-gen. No training recipe significantly outperforms either the untrained term-injection baseline or a rewriter-free concatenation on the deployment metric, except one Bonferroni-borderline NFCorpus cell (SFT-grounded GRPO). We also observe that the magnitude tracks the vocabulary gap (positive on wide- and medium-gap corpora, non-significant on narrow-gap) across our four anchor corpora and the out-of-sample medium-gap probe. This is a descriptive organizer on the three natural-corpus anchors, not a predictive partition rule. Positive-regime gains are robust within Qwen3. Llama-3.2 and Phi-3.5 collapse on FiQA and TC-gen due to rewrite-length failures, though Llama remains positive on the widest-gap corpus.

1 Introduction

LLM-based query rewriting for dense retrieval combines several pipeline components: an LLM that supplies domain vocabulary, a prompt-level injection of that vocabulary, an optional training stage that updates the rewriter on retrieval signal, and a retrieval-integration form that fuses rewritten and original queries. Reported gains on BEIR (Thakur et al., 2021) vary in sign across corpora and the literature does not isolate *which component* carries the effect (Zhu et al., 2025). Without that decomposition, proposed improvements cannot be cleanly attributed.

We ablate each component on four BEIR corpora at varying vocabulary-overlap intensities, two dense retrievers, and three rewriter families, with a held-out BioASQ (Krithara et al., 2022) medium-gap check. Our central finding is that the rewriting effect is primarily explained by *injected term content* rather than the rewriter, training recipe, or prompt.

Term injection is the principal factor (C1). Replacing extracted domain terms with count-matched random corpus words degrades retrieval sharply, significantly so on NFCorpus and TC-gen, isolating the extracted terms as the active ingredient rather than the injection mechanic.

Training adds limited gain (C2). No training recipe significantly exceeds the untrained term-injection baseline on $q_0 \oplus q_1$ fusion after multiplicity correction, except a Bonferroni-borderline NFCorpus cell (SFT-grounded GRPO).

Headroom lies in extractor quality (C3). Oracle extractors using gold-positive documents exceed the deployable recipe on every corpus, locating the ceiling in term selection rather than rewriter training.

Magnitude tracks vocabulary gap (C4). Rewriting helps on wide- and medium-gap corpora and is non-significant on narrow-gap across our four anchor corpora and one out-of-sample BioASQ point, consistent with the inconsistent signs in prior work. We treat the gap axis as a descriptive organizer on three natural-corpus anchors rather than a predictive partition rule.

2 Related Work

2.1 LLM expansion via pseudo-documents and corpus grounding

Pseudo-document expansion (Wang et al., 2023a; Gao et al., 2023; Mao et al., 2021) generates document-like context from an LLM's prior knowledge (Petroni et al., 2019). Query Expansion (Jagerman et al., 2023), CSQE (Lei et al., 2024), and the Generic Query Reformulation family (Wang et al., 2023b; Dhole and Agichtein, 2024; Jia et al., 2024) operate at finer granularities than pseudo-document injection. Jagerman et al. (2023) appends Flan-UL2 zero-shot/few-shot expansion text to the query for BM25 retrieval, with no fine-tuning. CSQE extracts pivotal sentences from a feedback-document set retrieved by an initial query pass. We instead represent the LLM's contribution as a discrete per-query term bank extracted from the raw query alone, separating the injected vocabulary from the rewriter that uses it. A separate LLM-for-IR direction synthesizes *training data for the retriever* (Bonifacio et al., 2022; Dai et al., 2022) or expands documents at index time with synthesized queries (Nogueira et al., 2019; Nogueira and Lin, 2019). Those modify the retriever or the index rather than the inference-time query and are complementary to the inference-time decomposition we study.

2.2 Corpus-level diagnostics for rewriting

The vocabulary gap is corpus-conditional. Independent describers of the same concept agree on the same word only $\approx 20\%$ of the time (Furnas et al., 1987), motivating retrieval-as-translation (Berger and Lafferty, 1999). The residual mismatch after a dense retriever still varies sharply by corpus. Prior diagnostics address this at two levels. *Within a corpus*, Query Performance Prediction (Cronen-Townsend et al., 2002; Hauff et al., 2008) and selective query expansion (Amati et al., 2004; Cronen-Townsend et al., 2004) predict per-query utility. *Across corpora*, Bigdeli et al. (2026) document that LLM-rewriting gains are inconsistent across BEIR, but leave open whether this variation can be summarized by a corpus-level statistic. We use the mean q–doc Jaccard (J) as a descriptive organizing variable across the five corpora we test, not a predictive partition rule. The closest diagnostic is NAQNR (Kotte, 2026), which studies per-query selective rewriting *within* a corpus. J organizes the sign *across* the corpora we test.

2.3 Trained rewriters and classical / dense pseudo-relevance feedback (PRF)

Trained rewriters with retrieval feedback. RaFe (Mao et al., 2024) trains a rewriting LLM with reranker feedback (Schulman et al., 2017; Rafailov et al., 2023; Ethayarajh et al., 2024). RL-QR (Cha et al., 2025) uses Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to train a query rewriter (Qwen3) with retrieval-based rewards computed from synthetic query–document pairs. Their untrained Qwen3 baseline degrades versus the raw query, which is compatible with our C1 claim that rewriting without injected terminology may fail to improve retrieval. Their RL gain over that no-injection floor is consistent with the policy learning to surface domain vocabulary into the query. Our comparison scopes the further question of whether training adds measurably over an injection-aware base+T baseline. Related but complementary directions blend multiple reformulations at retrieval time (Zhang et al., 2024; Wang et al., 2024) or jointly train expander and retriever (Yao et al., 2025). Our pipeline uses a single q_1 with the retriever held fixed. Table 1 summarizes the bank-level component differences from CSQE, RaFe, and NAQNR.

Classical and index-side approaches. Pre-LLM expansion methods such as Rocchio (Rocchio, 1971) and RM3 (Abdul-Jaleel et al., 2004; Lavrenko and Croft, 2001) derive expansion terms from pseudo-relevant documents and collection statistics (Carpineto and Romano, 2012). Dense pseudo-relevance feedback methods (Yu et al., 2021; Wang et al., 2023c; Zheng et al., 2020) generalize this idea by deriving expansion signals from learned dense representations of feedback documents rather than purely lexical statistics.

Retriever and index-side methods including DPR (Karpukhin et al., 2020), SPLADE (Formal et al., 2021b,a), Doc2Query (Nogueira et al., 2019), and Contriever (Izacard et al., 2021) address vocabulary mismatch through learned retrieval or document-side expansion.

Relationship to classical term-quality intuitions. Classical PRF work has observed that expansion-term quality matters for retrieval (Carpineto and Romano, 2012), but no prior study isolates this factor through controlled ablation. Rocchio (Rocchio, 1971) proposes a relevance-feedback vector formula without decomposing term source from in-

tegration method. RM3 (Abdul-Jaleel et al., 2004; Lavrenko and Croft, 2001) compares interpolation variants, not term-quality levels. Neither runs a random-term control, and both operate on lexical retrievers. Three features of the LLM pipeline make the question worth revisiting empirically: LLM rewriters generate novel terms from parametric knowledge rather than selecting from retrieved documents, so the term–method decomposition is no longer transparent; RL-trained rewriters (Mao et al., 2024; Cha et al., 2025) learn end-to-end policies that could plausibly discover term combinations a static bank cannot capture; and dense retrievers respond to distributional similarity, so term-level reasoning from lexical IR may not transfer. Our decomposition provides the first controlled test of whether these differences change the classical hierarchy.

Method	Bank source	Injection	Deploy
CSQE	fb docs^r	query str.	subst./fuse
RaFe	reranker rew.	trained rew.	fused
NAQNR	per-q predictor	routing	per-query
Ours	**LLM prior**^{o}	**prompt bank**	**gap-cond.**

Table 1: Component differences from CSQE, RaFe, and NAQNR. ^{r} retrieval pass at extraction time; ^{o} offline, retrieval-free.

3 Method

Overview. Our approach, **term injection**, has three stages. An offline LLM *extractor* emits a per-query domain-term bank from q_0 alone. The bank is inserted as a keyword list into a fixed prompt template to yield q_1, and the generated q_1 is fused with q_0 at retrieval time. Banks are precomputed once and reused across rewriters and retrievers.

Term extraction. The extractor is GPT-5-mini, prompted with per-corpus few-shot exemplars that map short queries to domain-vocabulary-rich rewrites. Given q alone, it emits a realistic and domain-enriched rewrite q^* on the four BEIR corpora. BioASQ OoS extractor variants are described with the OoS check. The per-query **term bank** is the set of word-types in q^* absent from q, under whitespace tokenization. Banks average 4–6 terms per query (6.1 NFCorpus, 5.4 TC-gen, 4.9 FiQA, 3.8 SciFact).

Extractor variants. The four extractors differ only in the source of the n terms: a GPT-5-mini rewrite of the query (*LLM*), the query's gold-positive documents (*Oracle*), the BM25 top-3 retrieved documents (*BM25-top-3*), or the corpus vocabulary (*Random*). n is the per-dataset LLM-bank mean (3–11 terms on NFCorpus). *LLM* takes the new word-types in a realistic rewrite q^*, the default extractor throughout, which we call the *deployable extractor*. *Oracle* picks the top-n terms by TF-IDF (Salton and Buckley, 1988) in the gold-positive documents, a label-using ceiling. *BM25-top-3* picks the n terms most common across the top-3 documents, a non-LLM baseline. Other non-LLM keyword baselines such as KeyBERT (Grootendorst, 2020) and YAKE (Campos et al., 2020) could substitute. We use BM25-top-3 for parameter-free comparison with the BM25 layer of the retrieval stack. *Random* samples n terms uniformly from corpus terms appearing in 3 documents to 60% of the collection, an uninformative-bank floor.

```
System: You are a query rewriter optimizing
for
document retrieval performance.
Input:
  query: {query}
  domain terms: {extracted_terms}
Rules (6 lines, see Appendix H):
  select / combine / recall-focused / . . .
Output: Rewritten query: {q1}
```

```
Example (NFCorpus, plain-721):
  query: BMAA
  extractor q*: beta methylamino L alanine
    BMAA neurotoxin cyanobacteria
  domain terms: alanine, beta, cyanobacteria,
    methylamino, neurotoxin
⇒ q1: BMAA methylamino neurotoxin
    cyanobacteria
```

Figure 1: The rewriter combines a per-query term bank into q_1 under six recall-focused rules. Full template in Appendix H.

Prompt template. The rewriter takes q_0 and the term bank as input and emits q_1. Figure 1 is the skeleton, with the full template in Appendix H. To test whether the lift comes from training-time exposure to terms or from inference-time access to them, we also run mismatched variants that supply the bank only at training or only at inference.

Retrieval pipeline. All experiments fuse the BGE-M3 dense and BM25 retrievers with reciprocal rank fusion (RRF) (Cormack et al., 2009). We report five retrieval modes:

- $q_{0,\text{dense}}$: baseline, no rewriting

- $q_{0,\text{fused}} = q_{0,\text{dense}} \oplus q_{0,\text{BM25}}$: baseline fused across the dense and BM25 retrievers with no rewriting (Table 5)
- $q_{1,\text{dense}}$: rewritten query, dense retriever only
- $q_{1,\text{fused}} = q_{1,\text{dense}} \oplus q_{1,\text{BM25}}$: rewritten query, both retrievers
- $q_0 \oplus q_1$ dense: baseline fused with rewritten query on the dense retriever (the **deployment metric**)

Throughout, $\oplus$ denotes RRF over two ranked lists. The operands are two retrievers (modes 2, 4) or two queries on the dense retriever (mode 5). The deployment Δ we report is the NDCG@10 change from $q_{0,\text{dense}}$ to $q_0 \oplus q_1$ dense, written "fused $- q_0$".

A notation summary collecting all shorthands is in Appendix I, Table 18.

4 Experimental Setup

Corpora. Four BEIR corpora span the gap axis (Table 2): NFCorpus (Boteva et al., 2016), FiQA (Maia et al., 2018), SciFact (Wadden et al., 2020), and TC-gen (Thakur et al., 2021).[1] The gap J is the stop-word filtered token Jaccard between q_0 and gold-positive token sets averaged over test queries (Appendix G). BioASQ Task 13B Golden (Krithara et al., 2022), restricted to Medical Subject Headings (MeSH) 2022 PubMed identifiers (PMIDs), is the out-of-sample (OoS) medium-gap probe.

Corpus	n_{test}	J	S	Regime
NFCorpus	323	0.0022	0.331	wide
FiQA	648	0.0321	0.683	medium
TC-gen	500	0.0288	0.931	narrow‡
SciFact	300	0.0404	0.850	narrow
BioASQ (OoS)	265	0.0119	n/a†	medium

Table 2: Anchor BEIR corpora along the Jaccard gap axis, with BioASQ as the OoS medium-gap probe. J = mean q_0–doc Jaccard, S = q_0 NDCG@10 baseline, *Regime* = descriptive corpus position. †BioASQ q_0 not comparable to BEIR. ‡TC-gen narrow by saturation, not J (Appendix F).

TC-gen contributes to the mechanism probes (the random-T paired comparison is untrained, so the Doc2Query label overlap does not apply) but is excluded from the gap-axis claim because its GRPO reward and the test qrels both come from Doc2Query (Appendix F). Its trained row in Table 3 is reported alongside the three natural corpora.

[1]TC-gen is the BEIR Doc2Query (Nogueira et al., 2019) pseudo-label corpus, not the 50-topic TREC-COVID benchmark (Voorhees et al., 2021).

Retrievers. We use BGE-M3 (Chen et al., 2024) as the first semantic retriever and jina-embeddings-v2-base (Günther et al., 2023) as a second probe. For lexical retrieval, we use BM25Okapi (Robertson and Zaragoza, 2009), held fixed across corpora. Dense and lexical scores are combined using RRF (encoder, BM25, and fusion parameters in Appendix J).

Rewriters. We evaluate four rewriters across two architectures. Qwen3-4B base and Qwen3-4B-Instruct-2507 (Yang et al., 2025) share an architecture and contrast base vs. Instruct training. Llama-3.2-3B-Instruct (Dubey et al., 2024) and Phi-3.5-mini-Instruct (Abdin et al., 2024) are the cross-architecture probes.

Training recipe. We train each rewriter with GRPO (Shao et al., 2024) and low-rank adaptation (LoRA) (Hu et al., 2022), initialized from either the base model or a supervised-fine-tuning (SFT) checkpoint. The reward is BGE-M3 retrieval NDCG of q_1 against per-corpus qrels (configuration in Appendix J). Terminology injection is applied at both training and evaluation time.

Statistical protocol. All Δ values are reported with paired-bootstrap 95% CIs (Berg-Kirkpatrick et al., 2012). **Bold** marks deltas whose CI excludes zero after Bonferroni correction within each claim family. Where we report two conditions as "not significantly different," the 95% CI of the paired Δ includes zero. This is absence of evidence for a difference. We reserve "equivalent" for cases where the two one-sided tests (TOST) procedure (Lakens et al., 2018) additionally rejects both one-sided nulls at a pre-specified margin, providing evidence of absence. Several non-significant results in this paper do *not* pass TOST (Appendix C). Full statistical configuration is detailed in Appendix C, including family structure, correction levels, equivalence margins, and a joint Holm m=12 corroboration (Holm, 1979).

5 Experiments

5.1 Corpora along the Jaccard gap axis

Rewriting retrieves complementary documents on wide-gap corpora. For each test query, we take the Jaccard between the BGE-M3 top-10 document IDs returned by q_0 and by q_1 (base+T). Call

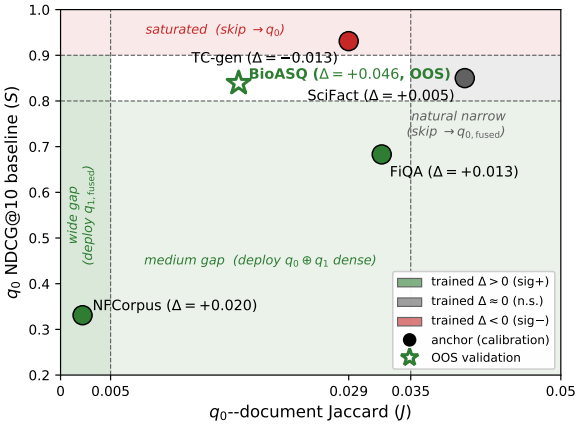

Figure 2: Sign of fused − q_0 tracks corpus position on the (J, S) plane (J = mean q_0–doc Jaccard, S = q_0 NDCG@10 baseline). Bands are corpus-relative descriptive labels, not predictive cutoffs.

Corpus	Recipe	base+T	rnd-T	trained
NFCorpus	SFT+GRPO	+0.008*	**−0.042**	**+0.022**
FiQA	SFT+GRPO	+0.007	−0.003	**+0.013**
TC-gen	GRPO	−0.009	**−0.031**	**−0.013**
SciFact	GRPO	+0.008	−0.009	+0.005

Table 3: Trained-rewriter Δ aligns with the q_0–doc Jaccard gap. Fused − q_0 NDCG@10 on $q_0 \oplus q_1$ dense, paired bootstrap (10k). **Bold**: nominal 95% CI excludes zero. Under Bonferroni m=4 within each condition family and joint Holm m=12, all bolded cells reject except TC-gen trained, which is nominal-only (Table 10). *NFCorpus base+T uses the first bank regeneration.

this the *top-10 overlap*. Top-10 overlap is low when the corpus gap is wide and high when it is narrow (Appendix E, Table 15). At the wide-gap extreme, NFCorpus's mean overlap is 0.27 and 9.6% of test queries return completely disjoint top-10 lists. At the narrow-gap extreme, SciFact has 0% disjoint queries. This explains the SciFact non-significant result: with overlap ≈0.6 the rewriter shifts only ≈40% of the top-10, and the K=10 operating point is near-ceiling in either direction.

Jaccard J as corpus-level summary statistic. The mean q–document Jaccard (J) captures this pattern at the corpus level: trained-rewriter Δ decreases with J across the three natural-corpus rows of Table 3 (BGE-M3 backbone). On these same three corpora J and the q_0-only baseline $S = q_{0,\text{dense}}$ NDCG@10 give a consistent ranking (NFCorpus < FiQA < SciFact), so the 3-point sign pattern is equally consistent with S as the organizer. Two pieces of evidence partially separate J from S. At the wide end, the NFCorpus random-T result rules out a pure-headroom reading of S: NFCorpus's q_0=0.331 sits well below ceiling, yet random-T drops q_1 NDCG@10 by 0.113, about a third of the baseline. The wide-end response depends on term content, not just on distance-from-ceiling. Our data do not resolve the FiQA–SciFact boundary. We therefore phrase C4 as "sign tracks J or a correlated property", not as a strict J-organizer claim. Figure 2 plots the four anchor corpora and the BioASQ OoS point in the (J, S) plane.

Descriptive, not predictive. Same-corpus robustness to retriever and rewriter family is addressed later via the retriever-swap and cross-family rewriter probes. New-corpus sign prediction is under-determined: any J-based cutoff is exact-fit on three corpora and sign-level LOO recovers NFCorpus only. The narrow-gap "skip" branch is extractor-conditional. Under the gold-positive oracle the SciFact result becomes significantly positive (+0.036, Table 8). "Skip" is therefore a deployable-extractor recommendation, not a claim that rewriting cannot help.

Out-of-sample medium-gap check on BioASQ. BioASQ (J=0.012; n_{docs}=35,123, 265/340 test queries retained) gives an out-of-sample medium-J point in the positive direction under the same extractor pipeline. Because BioASQ's collection differs from the BEIR corpora, its S is not on a comparable scale (Table 2 footnote), so the point confirms the medium-gap direction without separating J from S. Deployable-extractor base+T gives Δ = +0.017 [+0.003, +0.032], and the TF-IDF over gold-positives oracle opens Δ = +0.046 [+0.032, +0.060]. Both are reported uncorrected: BioASQ is a held-out validity check of the gap-axis pattern, not a within-family test, so the four-corpus m=4 partition does not extend to it. The reshuffling also shows up at the relevance level on wide-gap NFCorpus (Table 16): the $q_0 \cup q_1$ recall@10 ceiling sits well above either retriever alone, supplying the complementary coverage that RRF converts into NDCG@10 gain.

Estimating J without test set qrels. J requires gold relevance labels (qrels), which are unavailable at deployment. We consider two substitutes. (i) Computing J on the training set qrels (rather than the test set) recovers the correct regime label on the 4 natural corpora. The regime cutoffs were themselves fitted on these corpora, so this is a rank

	q_1 NDCG@10		Δ (rnd−ext)	
Corpus	base+T	rnd-T	q_1	fused
NFCorpus	0.331	0.218	**−0.113**	**−0.051**
FiQA	0.677	0.647	**−0.030**	−0.010
TC-gen	0.892	0.836	**−0.055**	**−0.022**
SciFact	0.859	0.832	**−0.028**	−0.017

Table 4: Random terms significantly degrade rewriting on the q_1-alone metric on every corpus. Paired bootstrap (10k); **bold** = nominal 95% CI excludes zero.

check, not a generalization claim. (ii) A fully label-free variant replaces the gold-positive documents with the BM25 top-10 retrievals (Appendix G, Table 17). The absolute J values compress, but rank order is approximately preserved, and 3 of 4 natural corpora (NFCorpus, FiQA, BioASQ) land in the correct regime while SciFact shifts to a neighboring regime. TC-gen is excluded from the J-proxy claim: its Doc2Query-generated labels place its regime outside J-determined classification.

5.2 The dominant pipeline component is the injected term bank

Three perturbations isolate the pipeline component carrying the principal effect: the term bank, the rewriter LLM and its training, and the retrieval integration form.

5.2.1 Term bank: content, presence, and source

We perturb the term bank along three dimensions: *content* (random-T), *presence* (empty term slot at inference), and *source* (extractor variants). The first two establish that term content carries the effect (C1). The third bounds the extractor ceiling (C3).

Content matters. Random term content significantly degrades fused NDCG@10 on NFCorpus and TC-gen (Table 4). Both conditions share the same Qwen3-4B base and a length-matched, prompt-fixed random extractor. Three confounds are controlled. (i) A frequency-matched bank on NFCorpus and TC-gen rules out a token-frequency artifact: freq-matched − uniform CIs cross zero on both. (ii) On TC-gen, near-ceiling cancels symmetrically in the paired Δ, so the random−extracted gap is not a ceiling effect. (iii) The Doc2Query label overlap does not apply to untrained comparisons.

Fused NDCG@10 drops significantly on NFCorpus and TC-gen. The TC-gen drop is shared by the untrained random-T (-0.031), so near-ceiling rather than Doc2Query-label overlap is the parsimonious reading (Appendix F). FiQA and SciFact are non-significant: the random-vs-extracted paired difference is small there. NFCorpus grounds C1: q_1 drops 0.113 NDCG@10 on a q_0 well below ceiling (the discriminator established by the gap-axis diagnostic). Training does not recover the gain: NFCorpus fused $\Delta = \mathbf{-0.044}$, overlapping prompt-only random-T and disjoint from trained-extracted.

No-bank at inference. With the term slot left empty at inference, the rewriting effect drops to q_0 on the positive-regime corpora. We compare each Δ to the per-corpus minimum detectable effect (MDE; values in Appendix C). FiQA falls from $+0.013$ to $+0.001$ ($\Delta = +0.012$, above MDE ± 0.011); NFCorpus from $+0.022$ to 0.000 ($\Delta = +0.022$, above MDE ± 0.016). TC-gen is within MDE in both conditions ($-0.010 \rightarrow -0.013$).[2] Domain knowledge does not live in the LoRA weights: without the term bank at inference, retrieval falls back to q_0-level. Output-level diagnostics on the same checkpoint (Appendix B) further rule out a retriever-gaming reading of the trained gain.

Extractor variants bound the term-bank ceiling. We run a term-bank ablation across the four extractor variants (current, oracle, random, and BM25-top-3) on all four corpora, holding rewriter / prompt / pipeline constant and length-matching the banks (mean ≈4–6 terms per query; per-cell numbers in Appendix A.3, Table 8; current LLM-extractor cell in Table 6). Ordering is consistent (oracle $\geq$ current $\geq$ BM25 $>$ random) on q_1 on every corpus, with BM25-extracted terms significantly underperforming LLM-prompted extraction on NFCorpus and FiQA q_1. **The oracle is significantly positive on both narrow-gap corpora** ($+0.022$ TC-gen, $+0.036$ SciFact fused), and the oracle−current headroom is significantly positive on every corpus. Narrow-gap is therefore "no currently available extractor is good enough," not "skip rewriting." These are upper bounds: the oracle uses gold positives, unavailable at deployment.

The GPT-5-mini extractor is non-deterministic: regenerating the NFCorpus term bank produced two versions, a first regeneration (base+T = $+0.008$, used in Table 3) and a second regeneration

[2] All three with-bank/no-bank pairs use SFT-grounded GRPO, the recipe carrying the NFCorpus increment in Table 11. For TC-gen this differs from Table 3, where the deployable recipe is GRPO-from-base.

(base+T = $+0.019$, used for training and the extractor ablation in Table 6). The 0.011 NDCG@10 gap between the two regenerations equals the NFCorpus MDE. All paired comparisons within Table 3 share the first regeneration, so within-table contrasts are unaffected.

5.2.2 Rewriter LLM and training add limited gain over base+T

The rewriter is not uniformly required. We define a rewriter-free $q_{1,\text{concat}} = q_0 + \text{extracted_terms}$ on the same test set term bank. Nominal 95% CIs of (concat $-$ trained) cross zero on every fused deployment corpus. NFCorpus q_1-alone is the only result where the nominal 95% CI excludes zero ($\Delta = -0.019$ $[-0.037, -0.001]$), but under Bonferroni correction ($m{=}4$ within metric or $m{=}8$ across all 8 contrasts) no result excludes zero (Table 7). TOST at $\delta{=}0.005$ (Appendix C) does not establish equivalence on any corpus. At the secondary MDE diagnostic, NFCorpus is bounded-worse (concat $<$ trained) while FiQA and TC-gen are bounded-close (Table 12). Per-query oracle $\max(\text{concat}, \text{trained})$ improves on the deployable recipe on NFCorpus / FiQA / TC-gen (Appendix C).

Training-recipe comparison. With the same injected term bank supplied to both the trained recipe and the base+T baseline on $q_0 \oplus q_1$ dense, three of four recipes (SFT-only, vanilla GRPO-from-base, rollout-variance-filtered GRPO) cross zero against base+T on every corpus. A mechanistic factor: across the four corpora, only 19–49% of training queries have non-degenerate across-rollout NDCG@10 spread (std > 0.05; per-corpus counts in Appendix J), so on the majority of queries the GRPO group-relative advantage collapses to zero and contributes no gradient signal. The fourth (SFT-grounded GRPO) is non-significant on FiQA / SciFact / TC-gen and shows a single NFCorpus-specific increment ($+0.013$ dense, $+0.018$ $q_{1,\text{fused}}$; Table 11). The outlier does not carry over to Qwen3-4B-Instruct ($\Delta = -0.001$, non-significant). Our non-significant finding answers the further question of whether training adds measurably over an injection-aware base+T baseline, rather than against the no-injection floor where RL-QR (Cha et al., 2025) reports its gains.

Reward–metric mismatch is not driving the non-significance. The q_1-alone reward is by design: training q_1 to maximize independent retrieval quality encourages complementary coverage under RRF, whereas a fused reward would incentivize q_1 to match q_0's ranking. Concretely, the training reward is per-rollout BGE-M3 NDCG@$K{=}100$ on q_1, so the deployment-metric non-significance could be a measurement artifact rather than absent gain. We test this by inspecting q_1-alone deltas (SFT-grounded GRPO $-$ base+T, paired bootstrap). Only FiQA shows the attenuation pattern: q_1-dense $+0.014$ (Bonferroni-borderline) versus fused $+0.005$. SciFact ($+0.001$) and TC-gen (-0.002) show no q_1-alone gain, and GRPO-from-base shows none on any corpus. Their non-significance is therefore genuine, not artifact-driven. NFCorpus SFT-grounded GRPO shows gains across both metrics, so the mismatch reading applies only to FiQA under SFT-grounded GRPO.

5.2.3 Retrieval integration form: deployment mode per-corpus

The integration form, whether and how q_1 enters retrieval, is the third pipeline component. Of the five retrieval modes defined in Method, the per-corpus deployable selection lands on different modes (Table 5): wide- and medium-gap corpora gain from $q_0 \oplus q_1$ dense fusion, while narrow-gap and saturated corpora are best served by skipping rewriting. Base+T is the deployable choice over trained recipes on the corpora we test (Table 11).

Corpus	Deployed mode	NDCG@10 (Δ vs $q_{0,\text{dense}}$)
FiQA	$q_0 \oplus q_1$ dense	0.690 ($+0.008$, n.s.)
NFCorpus	$q_0 \oplus q_1$ dense	0.339 ($+0.008$, n.s.)
SciFact	$q_{0,\text{fused}}$	0.876 (no rewriting)
TC-gen	$q_{0,\text{dense}}$	0.931 (no rewriting)

Table 5: Under the trained recipe, fusion is not significantly better than $q_{0,\text{dense}}$ on FiQA or NFCorpus. n.s.: 95% CI of Δ includes zero.

5.3 Classical and pseudo-document baselines do not match the cross-regime pattern

Pseudo-document injection. HyDE (Gao et al., 2023) and Query2Doc (Wang et al., 2023a) generate a passage-length context from the rewriter's prior. We evaluate at equal-weight RRF, matching our deployment setup. We do not sweep asymmetric per-corpus RRF weighting (Gao et al., 2023), so HyDE/Q2D are reported at default rather than at a configuration tuned with the same per-corpus protocol as RM3 below. This asymmetry is a known scope-bound on comparisons. Under this configuration the regime-conditional sign matches the C4

gap-axis pattern on 3/4 corpora. Per-corpus CIs in Appendix D, Table 13.

Classical PRF (standard RM3). Standard RM3 (Abdul-Jaleel et al., 2004; Lavrenko and Croft, 2001) uses the λ original-query mixture with per-corpus tuning over (λ, K, N) on the validation set (full description in Appendix D). It is significantly positive on NFCorpus on both BM25-only and dense+BM25 fusion contrasts (Bonferroni $m{=}4$ retains significance), but produces no significant effect on FiQA, SciFact, or TC-gen (Table 14). PRF sources expansion terms from the same retriever's BM25 top-k of q_0, so RM3 produces a significant gain in the wide-gap regime where BM25 already retrieves lexically relevant context for expansion. On medium-gap (FiQA) the LLM-extractor pipeline produces a significant gain (Table 3) while RM3 does not.

5.4 Robustness across retriever and rewriter family

5.4.1 Dense retriever swap

Re-evaluating all four corpora on Jina v2 base as a drop-in dense replacement (BM25 unchanged), no significant cell flips sign across the 12 contrasts (4 corpora $\times$ 3 conditions, including random-T). Two non-significant results oscillate in sign within MDE rather than flipping (FiQA base+T, SciFact random-T). Jina v2 amplifies effects in both directions (e.g., NFCorpus trained $+0.022 \rightarrow +0.029$; TC-gen random-T $-0.031 \rightarrow -0.073$).

5.4.2 Cross-family rewriter probe: asymmetric transfer

We test transfer to Llama-3.2-3B-Instruct and Phi-3.5-mini-Instruct, with Qwen3-4B-Instruct-2507 as a same-family positive control (Appendix B, Table 9). The negative regime carries over: TC-gen's significantly-negative direction and SciFact's non-significant result hold across all three Instruct rewriters. The positive-regime corpora do not: FiQA goes significantly negative on both Llama-Inst and Phi-Inst even after GRPO-from-base, and NFCorpus stays near zero. The discriminator is not the Instruct label: same-family Qwen3-Inst matches Qwen3-base on the deployment metric (max $|\Delta| = 0.008$), and trained Qwen-Inst preserves the four-corpus signed pattern. The discriminator is instead a *length-collapse* failure mode on FiQA rewrites. Phi additionally collapses without injection, showing the failure originates in the rewriter itself, not in injection (Appendix B). The positive-regime claims (Tables 3, 5) are therefore scoped to rewriters passing a length screen: median $\leq$ 20 and p90 $\leq$ 25 whitespace tokens (Appendix B). The screen operationalizes the pre-specified query-like output format from Appendix H (4–10 words preferred, 12 max), excluding rewriters whose outputs systematically violate the design spec for the realistic q^* format the rewriter is intended to emit.

6 Conclusion

Across four BEIR corpora and a BioASQ OoS test, we attribute the principal effect to terminology content rather than the generative rewriter, training recipe, or prompt: replacing extracted terms with random corpus words degrades retrieval sharply, confirming rather than overturning the term-quality hierarchy from classical query expansion. The benefit tracks the query–document vocabulary gap: positive on wide- and medium-gap corpora, non-significant on narrow-gap. Training adds detectable gain only on the wide-gap NFCorpus (SFT-grounded GRPO above base+T, Bonferroni-borderline). Elsewhere base+T with $q_0 \oplus q_1$ fusion matches trained within noise at lower training cost, and the only nominally negative trained cell (TC-gen, -0.013) does not reject under Bonferroni $m{=}4$ or joint Holm $m{=}12$. Regime assignment to a new corpus requires a labeled probe-set rather than J alone. Beyond the recipe, headroom is in extractor quality and per-query oracle gains. Cross-family transfer is limited by rewrite-length control: Llama-3.2 and Phi-3.5 length-collapse on FiQA and TC-gen, scoping the positive regime to the Qwen3 family in our probe rather than the corpus gap alone. The negative regime carries over across all families.

Limitations

Our evidence covers four BEIR corpora and a BioASQ probe (all English, BioASQ restricted to a 35K MeSH-2022 subset), two encoder-only dense retrievers, and four rewriters, leaving the gap-axis pattern untested on non-English corpora, late-interaction / sparse / cross-encoder retrievers, and pretrained-only different-family rewriter bases. The ordering rests on three natural anchor corpora plus BioASQ, so the regime cutoffs are under-determined, and we do not claim equivalence on the SciFact non-significant result. Two

design choices were held fixed: the GRPO reward is q_1-alone NDCG@K=100 while the deployment metric is fused NDCG@10 on $q_0 \oplus q_1$ dense, and HyDE/Query2Doc baselines use equal-weight RRF without per-corpus tuning. The deployable recipe depends on the proprietary GPT-5-mini extractor (one offline LLM call per query). The BM25-top-3 non-LLM variant (Table 8) bounds the LLM contribution from below on NFCorpus but cannot separate domain knowledge from pretraining overlap on these public corpora, and cross-run bank variance was measured only on NFCorpus, within the per-corpus MDE.

Ethical Considerations

All five corpora are research benchmarks released under their respective public licenses (see dataset repositories). We collect no new user data and conduct no human-subjects studies. Three corpora are biomedical (NFCorpus, SciFact, BioASQ). The term banks distill clinical vocabulary from GPT-5-mini and may inherit its training-data biases (over-representation of English-language clinical practice, under-representation of non-Western medical traditions). The banks support *retrieval* only and are not intended for patient-facing use or to substitute for clinical judgment. The trained rewriter and extractor banks will be released for reproducibility.

Acknowledgments

Per ACL's AI Writing Assistance policy, we disclose use of Anthropic's Claude, limited to language polishing of author-written prose, captions, and tables, structural and rhetorical suggestions on already-decided content (section ordering, caption format, terminology rules), and compliance auditing against venue rules. All research ideas, experimental design, analysis code, statistical procedures, numerical results, and final wording are the authors'. The model contributed no new research claims, and the authors verified all citations. Named-person acknowledgments are omitted for anonymous review and will be added at camera-ready.

References

Marah Abdin and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah S. Larkey, Xiaoyan Li, Mark D. Smucker, and Courtney Wade. 2004. UMass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*. Introduces the RM3 query-mixture variant of the Lavrenko–Croft relevance model.

Giambattista Amati, Claudio Carpineto, and Giovanni Romano. 2004. Query difficulty, robustness, and selective application of query expansion. In *Advances in Information Retrieval — 26th European Conference on IR Research (ECIR)*, pages 127–137.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229.

Amin Bigdeli, Radin Hamidi Rad, Hai Son Le, Mert Incesu, Negar Arabzadeh, Charles LA Clarke, and Ebrahim Bagheri. 2026. A reproducibility study of llm-based query reformulation. *arXiv preprint arXiv:2604.27421*.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. InPars: Data augmentation for information retrieval using large language models. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2387–2392.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval — 38th European Conference on IR Research (ECIR)*, pages 716–722.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Mário Jorge, Célia Nunes, and Adam Jatowt. 2020. YAKE! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1:1–1:50.

Sungguk Cha, DongWook Kim, Taeseung Hahn, Mintae Kim, Youngsub Han, and Byoung-Ki Jeon. 2025. Annotation-free reinforcement learning query rewriting via verifiable search reward. *arXiv preprint arXiv:2507.23242*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics.

Gordon V. Cormack, Charles L.A. Clarke, and Stefan Buettcher. 2009. 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 (SIGIR)*, pages 758–759.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2002. Predicting query performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 299–306.

Steve Cronen-Townsend, Yun Zhou, and W. Bruce Croft. 2004. A framework for selective query expansion. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM)*, pages 236–237.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint arXiv:2209.11755*.

Kaustubh D. Dhole and Eugene Agichtein. 2024. GenQREnsemble: Zero-shot LLM ensemble prompting for generative query reformulation. In *Advances in Information Retrieval — 46th European Conference on Information Retrieval (ECIR)*, pages 326–335.

Abhimanyu Dubey and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. KTO: Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. 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.

George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Maarten Grootendorst. 2020. KeyBERT: Minimal keyword extraction with BERT.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, and 1 others. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *arXiv preprint arXiv:2310.19923*.

Claudia Hauff, Djoerd Hiemstra, and Franciska de Jong. 2008. A survey of pre-retrieval query performance predictors. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1419–1420.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. In *Proceedings of the 1st Workshop on Generative Information Retrieval (Gen-IR), SIGIR*.

Pengyue Jia, Yiding Liu, Xiangyu Zhao, Xiaopeng Li, Changying Hao, Shuaiqiang Wang, and Dawei Yin. 2024. MILL: Mutual verification with large language models for zero-shot query expansion. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Varun Kotte. 2026. Not all queries need rewriting: When prompt-only LLM refinement helps and hurts dense retrieval. *arXiv preprint arXiv:2603.13301*.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2022. BioASQ-QA: A manually curated corpus for biomedical question answering. *bioRxiv*.

Daniël Lakens, Anne M. Scheel, and Peder M. Isager. 2018. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269.

Victor Lavrenko and W. Bruce Croft. 2001. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 120–127.

Yibin Lei, Yu Cao, Tianyi Zhou, Tao Shen, and Andrew Yates. 2024. Corpus-steered query expansion with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL): Short Papers*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of The Web Conference 2018 (WWW)*, pages 1941–1942.

Shengyu Mao, Yong Jiang, Boli Chen, Xiao Li, Peng Wang, Xinyu Wang, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2024. RaFe: Ranking feedback improves query rewriting for RAG. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 884–901, Miami, Florida, USA.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to docTTTTTquery. Online preprint.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems (NeurIPS)*.

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

Joseph John Rocchio. 1971. Relevance feedback in information retrieval. In Gerard Salton, editor, *The SMART Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice-Hall.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*. Introduces Group Relative Policy Optimization (GRPO).

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. TREC-COVID: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1):1–12.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Cheng, Tuo Zhao, and Jing Gao. 2024. BlendFilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2023b. Generative query reformulation for effective adhoc search. *arXiv preprint arXiv:2308.00415*.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023c. ColBERT-PRF: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, 17(1):1–39.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Sijia Yao, Pengcheng Huang, Zhenghao Liu, Yu Gu, Yukun Yan, Shi Yu, and Ge Yu. 2025. ExpandR: Teaching dense retrievers beyond queries with LLM guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19036–19054, Suzhou, China.

HongChien Yu, Chenyan Xiong, and Jamie Callan. 2021. Improving query representations for dense retrieval with pseudo relevance feedback. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 3592–3596.

Le Zhang, Yihong Wu, Qian Yang, and Jian-Yun Nie. 2024. Exploring the best practices of query expansion with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. BERT-QE: Contextualized query expansion for document re-ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4718–4728.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zheng Liu, Zhicheng Dou, and Ji-Rong Wen. 2025. Large language models for information retrieval: A survey. *ACM Transactions on Information Systems*, 44(1):1–54.

Appendices

A Ablation Tables and Oracle Bounds

A.1 Per-corpus oracle bounds

Table 6 reports two ceilings per corpus: extractor-quality (oracle vs current bank) and rewrite-vs-skip ($\max(q_0, q_1) - q_0$). The $\max(\text{concat}, \text{trained})$ oracle is in Appendix C.

Corpus	Term-bank current	Term-bank headroom	Per-query oracle
NFCorpus	+0.019	**+0.043**	**+0.059**
FiQA	+0.016	**+0.021**	**+0.043**
TC-gen	−0.012	**+0.034**	**+0.022**
SciFact	+0.006	**+0.029**	**+0.028**

Table 6: Both oracles exceed the deployable recipe on every corpus. *Current* = deployable vs q_0; *headroom* = oracle − current; *per-query* = $\max(q_0, q_1) - q_0$; **bold** = 95% CI excludes zero (paired bootstrap, 10k). NFCorpus *current* uses the second bank regeneration, distinct from Tab. 3 base+T's first regeneration (+0.008).

A.2 Concat baseline vs trained rewriter

Table 7 gives the full concat-vs-trained comparison. TOST analysis is in Table 12.

Corpus	metric	concat	trained	Δ
NFCorpus	q_1	0.331	0.349	**−0.019**
NFCorpus	fused	0.342	0.352	−0.010
FiQA	q_1	0.686	0.691	−0.005
FiQA	fused	0.694	0.695	−0.001
TC-gen	q_1	0.899	0.889	+0.010
TC-gen	fused	0.925	0.922	+0.003
SciFact	q_1	0.864	0.855	+0.008
SciFact	fused	0.862	0.855	+0.007

Table 7: Rewriter-free concat is not significantly different from the trained rewriter on any fused deployment cell. Δ = concat − trained; **bold** = nominal 95% CI excludes zero (paired bootstrap, 10k).

A.3 Extractor variants

Table 8 gives the full four-way extractor comparison. The current LLM extractor reference is the *current* column of Table 6.

Corpus	oracle	random	BM25
NFCorpus	**+0.062**	**−0.038**	**−0.016**
FiQA	**+0.037**	−0.006	−0.004
TC-gen	**+0.022**	**−0.020**	−0.009
SciFact	**+0.036**	−0.003	+0.006

Table 8: Oracle terms are significantly positive on every corpus. Random terms degrade NFCorpus and TC-gen. Cells are fused NDCG@10 Δ vs q_0. **Bold** = 95% CI excludes zero (Bonferroni m=4 within each variant family; paired bootstrap, 10k).

B Cross-Family Rewriter Probe

We re-evaluated the four-corpus diagnostic on Llama-3.2-3B-Instruct and Phi-3.5-mini-Instruct (3B / 3.8B params) in three conditions: no-injection control, prompt-only with extracted terms, and GRPO-from-base trained. The Qwen3-4B *trained* column uses GRPO-from-base for matched comparison. Recipe substitution preserves the diagnostic sign on all four corpora. The two recipes agree within ±0.005 on FiQA, SciFact, and TC-gen, but differ by $|\Delta|$=0.014 on NFCorpus (the SFT-grounded outlier, Tab. 11). The cross-family NFCorpus comparison therefore uses the GRPO-from-base recipe (+0.008, Tab. 9) rather than the deployable SFT-grounded +0.022 (Tab. 3). Bonferroni m=4 throughout (Table 9).

Corpus	Diagnostic	Qwen3-4B	Llama-Inst	Phi-Inst
Base + extracted terms (fused − q_0)				
NFCorpus	positive	+0.008	+0.002	−0.013
FiQA	positive	+0.007	**−0.054**	**−0.039**
TC-gen	negative	−0.009	**−0.100**	**−0.111**
SciFact	no effect	+0.008	−0.010	−0.012
Trained rewriter (GRPO-from-base; fused − q_0)				
NFCorpus	positive	+0.008	+0.008	−0.012
FiQA	positive	+0.008	**−0.031**	**−0.022**
TC-gen	negative	−0.011	**−0.063**	**−0.086**
SciFact	no effect	+0.005	−0.008	−0.012
No-injection control (Phi-3.5; fused − q_0)				
NFCorpus	positive	n/a	n/a	**−0.013**
FiQA	positive	n/a	n/a	**−0.059**
TC-gen	negative	n/a	n/a	**−0.060**
SciFact	no effect	n/a	n/a	**−0.018**

Table 9: Llama-3.2 and Phi-3.5 base+T degrade FiQA and TC-gen. GRPO-from-base narrows but does not flip negative corpora. Phi-3.5 without injection is sig-negative on every corpus, isolating chat-collapse as upstream of injection. Fused − q_0 NDCG@10; **bold** = 95% CI excludes zero.

Post-training recovery across both families. GRPO-from-base narrows the FiQA / TC-gen gap (where base+T degraded most), but trained Instruct rewriters remain significantly worse than skipping rewriting on those corpora (Table 9).

No-injection control rules out the bank as the cause. Phi-3.5 produces significantly negative

fused $-$ q_0 on **every corpus** even without injection (Table 9, third block): Instruct training produces bad rewrites independent of term-bank content. Injection on Phi is non-significant on FiQA (q_1 inj $-$ no-inj $= +0.005$) and NFCorpus ($+0.004$), and actively *hurts* on SciFact (-0.032) and TC-gen (-0.178).

Same-family Instruct positive control. Swapping Qwen3-4B base for Qwen3-4B-Instruct-2507 isolates the Instruct-tuning factor. Qwen3-Inst + injection is non-significantly different from Qwen3-base + injection on $q_0 \oplus q_1$ dense across the four corpora (max $|\Delta| = 0.008$). Across three trained recipes (GRPO-from-base, SFT-only, GRPO-from-SFT-with-injection), trained $-$ base+T is non-significant on 11/12 results. The SFT-grounded lineage producing the Qwen3-base NFCorpus outlier is among these non-significant cells.

Output-collapse measurements. We measure two output features on FiQA test rewrites (n=648), each rewriter's raw output minus the `<think>` block. (i) **Rewrite length** (whitespace tokens; median / p90): Qwen3-base 13 / 20; Qwen3-Inst 14 / 20; Llama-3.2-3B-Inst 15 / 49; Phi-3.5-mini-Inst 37 / 45. Unrecovered families separate on the upper tail (p90). Phi additionally on the median. (ii) **BM25 top-10 Jaccard with the Qwen3-base rewrite** (mean over 648 paired queries): Qwen3-Inst 0.35; Llama-Inst 0.12; Phi-Inst 0.07. Qualitatively, Llama-Inst outputs frequently trail off into meta-commentary (e.g. "*This revised query incorporates key terms like* ...'), absent from Qwen3 rewrites.

Trained Qwen3-base reward-hacking screen. GRPO with KL penalty β=0.1 could in principle still reward-hack the BGE-M3 retriever (lexically degenerate but retrieval-scoring outputs). We verify this does not occur. We apply the length + Jaccard screen to the trained Qwen3-base rewriter (SFT-grounded GRPO; the checkpoint behind Table 3) on NFCorpus (n=323) and FiQA (n=648). Rewrite length (median / p90 / max whitespace tokens): NFCorpus $8/14/21$ vs base+T $8/14/28$; FiQA $12/18/30$ vs base+T $13/20/38$. Mean BM25 top-10 Jaccard with base+T retrievals (paired bootstrap, 1000 resamples): NFCorpus 0.282 $[0.258, 0.306]$; FiQA 0.319 $[0.299, 0.343]$. Both corpora pass the median$\leq$20, p90$\leq$25 screen. The trained max is *lower* than base+T's on both. The Jaccard sits between the recovered same-family point (0.35 Qwen-Inst) and the collapsed cross-family points (0.12 Llama-Inst, 0.07 Phi-Inst). No length-collapse or diversity-collapse: the trained rewrites are minor refinements of base+T.

C Statistical Protocol Detail

Family-wise correction rationale. The 12 cells of Table 3 (3 conditions $\times$ 4 corpora) are partitioned into three Bonferroni families *by inferential role*, not by data-driven cell selection: (i) *trained* cells, testing the hypothesis that the trained rewriter significantly outperforms q_0 (m=4); (ii) *base+T* cells, testing whether injection produces a detectable effect (m=4); (iii) *random-T* cells, testing whether content matters (m=4). These address distinct null hypotheses. Pooling into a single m=12 correction over-corrects by treating mechanism diagnostics as interchangeable with the deployment claim. Each robustness probe (Jina v2 backbone) is its own family at m=4, α=0.0125.

Joint Holm m=12 corroboration. As a partition-free fallback we report all 12 cells under a single Holm–Bonferroni correction (Holm, 1979). All four cells that reject under the three-family m=4 partition (NFCorpus random-T, NFCorpus trained, FiQA trained, TC-gen random-T) also reject under joint Holm m=12 at α=0.05 (Table 10), so the partition choice is not essential to the signed conclusions of Table 3. This corroboration covers the 12 mechanism contrasts against q_0. The trained-vs-base+T comparisons are a separate m=4 family in Table 11.

Minimum detectable effect. Per-corpus 95%-CI half-width / 80%-power MDE on fused $-$ q_0: FiQA $\pm 0.008/\pm 0.011$; TC-gen $\pm 0.010/\pm 0.014$; SciFact $\pm 0.010/\pm 0.015$; NFCorpus $\pm 0.011/\pm 0.016$.

Practical-equivalence margin. We test practical equivalence at δ=0.005 NDCG@10 (Lakens et al., 2018) as the stricter margin (one NDCG@10 percentage point as a deployment-resolution threshold, since no published IR-evaluation study we are aware of fixes this value, making it a design choice rather than an external standard). As a secondary diagnostic we report TOST at the per-corpus MDE, the smallest delta the paired-bootstrap design can reliably resolve at our query counts, which scales with sample size. Neither margin establishes absolute practical equivalence. Pass/fail is reported

Corpus	Cond.	Δ	95% CI	p_{raw}	p_{Holm}
NFCorpus	base+T	$+0.008$	$[-0.003, +0.020]$	0.148	0.740
NFCorpus	rnd-T	$\mathbf{-0.042}$	$[-0.057, -0.028]$	<0.001	**0.001**
NFCorpus	trained	$\mathbf{+0.022}$	$[+0.011, +0.033]$	<0.001	**0.001**
FiQA	base+T	$+0.007$	$[-0.003, +0.016]$	0.162	0.740
FiQA	rnd-T	-0.003	$[-0.012, +0.006]$	0.443	0.740
FiQA	trained	$\mathbf{+0.013}$	$[+0.006, +0.021]$	<0.001	**0.001**
TC-gen	base+T	-0.009	$[-0.018, +0.001]$	0.071	0.484
TC-gen	rnd-T	$\mathbf{-0.031}$	$[-0.044, -0.018]$	<0.001	**0.001**
TC-gen	trained	-0.013	$[-0.022, -0.002]$	0.018	0.146
SciFact	base+T	$+0.008$	$[-0.003, +0.018]$	0.156	0.740
SciFact	rnd-T	-0.009	$[-0.020, +0.001]$	0.069	0.484
SciFact	trained	$+0.005$	$[-0.005, +0.015]$	0.359	0.740

Table 10: Holm m=12 corroborates the four cells that reject under the family m=4 partition. **Bold** = reject at α=0.05 under Holm.

Corpus	Δ	95% CI
NFCorpus	$\mathbf{+0.013}^{\circ,\ddagger}$	$[+0.002, +0.024]$
FiQA	$+0.005$	$[-0.005, +0.014]$
TC-gen	$+0.001$	$[-0.009, +0.011]$
SciFact	-0.003	$[-0.010, +0.002]$

Table 11: SFT-grounded GRPO $-$ base+T on $q_0 \oplus q_1$ dense fused NDCG@10 (paired bootstrap, 10k). **Bold**: 95% CI excludes zero. °Bonferroni m=4 borderline. ‡On $q_{1,\text{fused}}$, NFCorpus is $+0.018$ $[+0.001, +0.035]$ (also borderline).

at each separately. MDE-TOST passage is expected given non-significance and is reported only for completeness, not as independent evidence of equivalence.

Bootstrap-TOST table. Paired-bootstrap symmetric 90% CIs of Δ = concat $-$ trained (10k resamples). The trained model per row is the strongest recipe for that corpus (SFT-grounded GRPO on FiQA/NFCorpus, GRPO-from-base on SciFact/TC-gen). The GRPO-from-base recipe deployed in Table 5 is non-significant vs q_0 on FiQA/NFCorpus, so concat-vs-trained would be at least as bounded-close under the deployed recipe. Per-query oracle gain $\max(\text{concat}, \text{trained})$ is $+0.023$ NFCorpus / $+0.026$ FiQA fused: the MDE-diagnostic indistinguishability is a statement about average performance only.

D Pseudo-Document and Classical PRF Baselines

HyDE / Query2Doc. On positive-regime corpora the generated passage emits non-target vocabulary that contaminates dense-only retrieval (FiQA dense -0.046 / -0.025), but q_0-fusion restores both methods to non-significance. On high-baseline narrow-gap (TC-gen) fusion remains significantly negative: the baseline cannot absorb the pseudo-passage's noise. SciFact's claim is that abstract pseudo-passage is closer to the corpus distribution and yields small significantly-positive fusion effects at the edge of the diagnostic's narrow-gap band.

Standard RM3 (Lavrenko relevance-model PRF). We implement standard RM3 (Abdul-Jaleel et al., 2004; Lavrenko and Croft, 2001) with Laplace-smoothed document language models, stop-word and query-token exclusion, and max-normalized BM25 score interpolation. We sweep $(\lambda, K, N) \in \{0.3, 0.5, 0.7\} \times \{10, 20, 50\} \times \{20, 40\}$ on the validation set (capped at 200 queries for FiQA/TC-gen) and report the test set under the best validation config. Selected (λ, K, N): NFCorpus $(0.5, 20, 20)$, FiQA $(0.5, 20, 40)$, SciFact $(0.3, 20, 40)$, TC-gen $(0.5, 50, 20)$.

E q_0 vs q_1 Retrieved-Pool Overlap

Tables 15 and 16 report the full per-corpus pool overlap and recall-union statistics.

F TC-gen Trained Result Readings

The TC-gen trained result (-0.013 fused $-$ q_0, Table 3) admits two non-exclusive readings that the design does not separate. The random-T result privileges the near-ceiling reading as the parsimonious explanation.

Reading 1 – near-ceiling effect. TC-gen's q_0 is near-ceiling on the fused metric (S=0.931, Table 2), so the fusion is too near ceiling to absorb rewrite noise. The prompt-only random-T result is significantly negative (-0.031, Table 3): noise on a near-ceiling q_0 subtracts NDCG independently of training or label source.

Corpus	Δ	90% CI	MDE	@MDE	@0.005
NFCorpus	-0.010	$[-0.019, -0.001]$	n/a	fails*	fails*
FiQA	-0.001	$[-0.009, +0.006]$	0.0107	passes	fails
TC-gen§	$+0.003$	$[-0.006, +0.012]$	0.0142	passes	fails
SciFact	$+0.007$	$[-0.002, +0.017]$	0.0148	fails	fails

Table 12: No corpus passes TOST at δ=0.005; FiQA and TC-gen pass at the secondary MDE. (concat $-$ trained) fused NDCG@10, paired bootstrap (10k); * bounded-worse (90% CI < 0); § TC-gen pass confounded by shared Doc2Query labels.

Variant	NFCorpus	FiQA	TC-gen	SciFact
HyDE dense	-0.012	$\mathbf{-0.046}$	$\mathbf{-0.158}$	$+0.013$
HyDE $q_0 \oplus$ d.	$+0.013$	-0.002	$\mathbf{-0.052}$	$\mathbf{+0.019}$
Query2Doc dense	$\mathbf{-0.026}$	$\mathbf{-0.025}$	$\mathbf{-0.106}$	$\mathbf{+0.046}$
Query2Doc $q_0 \oplus$ d.	$+0.003$	$+0.006$	$\mathbf{-0.027}$	$\mathbf{+0.024}$

Table 13: Fused HyDE and Query2Doc match the gap-axis per-corpus sign on 3/4 corpora. "$q_0 \oplus$ d." = RRF of q_0 and pseudo-document on the dense retriever; **bold** = 95% CI excludes zero (paired bootstrap, 10k).

Reading 2 – shared Doc2Query labels. TC-gen's qrels are Doc2Query pseudo-labels (Nogueira et al., 2019), and the GRPO reward retrieves against those same qrels, so train labels, test labels, and reward all derive from the same Doc2Query process. A rewriter trained on this reward inherits whatever idiosyncrasies the Doc2Query generator carries. The trained number could therefore reflect learning to fit those specific labels rather than real retrieval improvement.

G J-Proxy Validation Detail

Mean query–document Jaccard J is defined as

$$J = \frac{1}{|Q_{\text{test}}|} \sum_{q \in Q_{\text{test}}} \frac{|W(q) \cap W(D_q^+)|}{|W(q) \cup W(D_q^+)|}, \quad (1)$$

where $W(\cdot)$ is the lowercased stop-word filtered token set (Bird and Loper, 2004) and D_q^+ is q's full gold-positive token set. The four corpora + BioASQ are scored under three estimators that vary D_q^+. *test-gold* = test queries $\times$ test-gold positives; *train-gold* = train queries $\times$ train-gold positives; *BM25-top-10* = test queries $\times$ BM25-top-10 docs as pseudo-positives (no qrels needed). All three share the same tokenizer (NLTK English stop-word list, lowercase, punctuation strip) and formula. Only D_q^+ differs. This tokenization is distinct from the whitespace tokenization used for term-bank derivation in Method.

H Prompt Templates

Rewriter prompt (term-injection variant). The literal template used at training and evaluation, with placeholders `{query}` and `{extracted_terms}` populated per query:

```
You are a query rewriter optimizing
for document retrieval performance.
Your goal is to rewrite the query to
maximize relevant document retrieval
by strategically selecting and
incorporating the most effective
domain terms.

query: {query}
domain terms: {extracted_terms}

Rules:
- Select terms from the available list
that would appear in relevant documents
- Use all, some, or none of the provided
terms based on retrieval effectiveness
- Combine terms naturally with the
original query
- Focus on recall - retrieving more
relevant documents is better than
perfect precision
- Consider which terms experts would use
in target documents
- Output only the rewritten query, no
labels or prefixes

Rewritten query:
```

Realistic-generation prompt (GPT-5-mini). The offline extractor produces a "realistic" rewrite q^* from which the per-query term bank is derived (set-difference against q_0). The extractor runs in a query-only mode (no retrieved documents passed to the LLM), seeing only q_0 plus per-corpus few-shot exemplars. NFCorpus, SciFact, and TC-gen use analogous per-corpus exemplars sampled from each training set. The FiQA prompt is shown below:

```
Goal: Rewrite the Original Query into
one realistic, improved search query for
a real retrieval setting.

Hard constraints (all must hold):
1. Preserve the original user intent
```

Corpus	vs $q_{0,\text{BM25}}$	vs fused $q_{0,\text{dense}} \oplus q_{0,\text{BM25}}$
NFCorpus	$+$**0.017** [$+0.007$, $+0.026$]	$+$**0.021** [$+0.009$, $+0.033$]
FiQA	-0.003 [-0.011, $+0.005$]	-0.004 [-0.016, $+0.008$]
TC-gen	-0.008 [-0.019, $+0.002$]	$+0.000$ [-0.013, $+0.014$]
SciFact	-0.003 [-0.016, $+0.011$]	$+0.002$ [-0.019, $+0.023$]

Table 14: RM3 is significantly positive on NFCorpus only. Left: Δ vs q_0 BM25; right: Δ vs RRF($q_{0,\text{dense}}, q_{0,\text{BM25}}$) with expanded BM25 substituted. Paired bootstrap, 10k; **bold** = 95% CI excludes zero; Bonferroni m=4 retains NFCorpus.

		K=10		K=5		
Corpus	N	Jacc	%disj	Jacc	%disj	J
NFCorpus	323	0.267	9.6%	0.270	19.5%	0.0022
FiQA	648	0.551	0.2%	0.561	1.4%	0.0321
TC-gen	500	0.426	0.0%	0.435	1.0%	0.0288
SciFact	300	0.600	0.0%	0.602	0.0%	0.0404

Table 15: Top-K pool overlap between q_0 and q_1 increases monotonically with the corpus gap axis. %disj = fraction of test queries with a completely disjoint top-K list.

Corpus	R_{q_0}	R_{q_1}	$R_{q_0 \cup q_1}$	Δ_{union}	%rel-disj
NFCorpus	0.155	0.161	0.194	+0.033	49.8%
FiQA	0.750	0.743	0.786	+0.036	10.8%
TC-gen	0.978	0.962	0.986	+0.008	3.2%
SciFact	0.921	0.927	0.941	+0.014	3.6%

Table 16: Only wide-gap NFCorpus shows a meaningful $q_0 \cup q_1$ recall gain over either query alone. $\Delta_{\text{union}} = R_{q_0 \cup q_1} - \max(R_{q_0}, R_{q_1})$; %rel-disj = fraction of relevant docs found by only q_0 or only q_1.

Corpus	test	train	BM25-10	regime
NFCorpus	0.0022	0.0024	0.0041	wide
FiQA	0.0321	0.0328	0.0114	medium
TC-gen	0.0288	0.0287	0.0044	narrow
SciFact	0.0404	0.0417	0.0100	narrow
BioASQ	0.0119	0.0162	0.0069	medium

Table 17: Train-gold J matches the test-gold regime label on the 4 natural corpora, and BM25-top-10 preserves the regime on 3 of 4. The check is on computing J from non-test signals, not on extending the fitted cutoffs to unseen corpora. TC-gen is shown for context but is a synthetic Doc2Query probe with regime determined by saturation rather than J.

```
exactly.
2. Output a query-like string, not a
sentence.
3. Length: 4-10 words preferred, 12
max.
4. Use only high-signal terms (domain
terms, entities).
5. Ignore one-off specifics that look
incidental.
6. Do not introduce side-topic concepts.
7. Do not copy long spans (>5
consecutive words).
8. No explanations, quotes, or labels.

Few-shot FiQA Examples:
Example 1.
Original Query: business debt and taxes
Output Query: business debt to equity
ratio s corp tax liability

Example 2.
Original Query: tax on investment
Output Query: capital gains tax 1031
exchange investment requirements

Example 3.
Original Query: workplace retirement
plans
Output Query: solo 401k contribution
limits self employed retirement plan

Now solve the actual case:
Original Query: {original_query}

Output: exactly one line containing only
the rewritten query.
```

I Notation Summary

Sym.	Meaning	Sym.	Meaning
q_0	original query	q_1	rewritten query
$\oplus$	RRF fusion	base+T	base + ext. terms
rnd-T	base + rnd. terms	trained	GRPO/SFT + terms
concat	q_0+terms, no LLM	J	q_0–doc Jaccard
S	q_0 NDCG@10	MDE	min. det. effect
fused$-q_0$	NDCG@10($q_0 \oplus q_1$) $-$ NDCG@10(q_0)		

Table 18: Notation and condition shorthands.

J Training Hyperparameters

Table 19 lists all training, decoding, and retrieval configurations used throughout the experiments.

Rollout-variance-filtered GRPO. This recipe initializes from the SFT-with-injection model and trains on a rollout-variance-filtered training set. For each training query we sample K=16 pilot rollouts from the SFT-grounded policy at T=1.4 (matching GRPO training temperature) and compute their NDCG@10 against the per-corpus qrels. Queries

Field	GRPO-from-base	SFT-with-injection
Base model	Qwen3-4B	Qwen3-4B
Epochs	1	3
Learning rate	1×10^{-6}	3×10^{-6}
LR schedule	cosine	cosine
Warmup ratio	0.20	0.10
Weight decay	0.015	0.01
Per-device batch	8	16
Grad-accum steps	2	2
Effective batch	16	32
Precision	bf16	bf16
Grad checkpointing	yes	yes
DeepSpeed	ZeRO-2	ZeRO-2
Optimizer	AdamW	AdamW
LoRA r/α (FiQA)	16 / 32	16 / 32
LoRA r/α (NFCorpus, SciFact, TC-gen)	8 / 16	8 / 16
LoRA dropout	0.1	0.1
LoRA target modules	q,k,v,o,gate,up,down	q,k,v,o,gate,up,down
LoRA bias	none	none
GRPO group size G	16	–
GRPO temperature T	1.4	–
GRPO max_completion_length	64	–
GRPO max_prompt_length	256	–
GRPO KL coefficient β	0.1	–
GRPO clip ratio ϵ	0.2	–
SFT max input length	–	512
SFT max target length	–	64
Best-model metric	eval q_1 NDCG@10	eval q_1 NDCG@10
Early-stopping patience	8 evals	n/a
Reward	per-rollout NDCG@K=100 of BGE-M3 retrieval on q_1	
Seed	42	
Hardware		
GPUs	4× NVIDIA A100	
Approx. training time	2–4 hours per corpus (GRPO-from-base)	
Decoding		
do_sample	false	
repetition_penalty	1.2	
enable_thinking	false	
Retrieval		
Dense retriever	BGE-M3 (1024-d, multilingual)	
Dense retriever (probe)	jina-embeddings-v2-base (768-d)	
Lexical retriever	BM25Okapi (k_1=1.5, b=0.75)	
RRF k / weights	60 / equal (dense+BM25)	
Extractor API (GPT-5-mini)		
temperature	1.0	
top_p	1.0	
reasoning_effort	minimal	
verbosity	low	

Table 19: Training, decoding, retrieval, and extractor API configuration.

are retained iff the across-rollout NDCG@10 standard deviation exceeds 0.05. Per-corpus retention: NFCorpus 336 / 1630, FiQA 1037 / 5500, SciFact 213 / 920, TC-gen 1942 / 4000. Hyperparameters otherwise match the GRPO column of Table 19, except learning rate 5×10^{-7} and warmup ratio 0.4 to compensate for the smaller training set.