NOT ALL QUERIES NEED REWRITING: WHEN PROMPT-ONLY LLM REFINEMENT HELPS AND HURTS DENSE RETRIEVAL

Varun Kotte
Adobe
vkotte@adobe.com

ABSTRACT

Prompt-only, single-step LLM query rewriting (i.e., a single rewrite generated from the query alone, without retrieval feedback) is commonly deployed in production RAG pipelines, but its impact on dense retrieval is poorly understood. We conduct a systematic empirical study across three BEIR benchmarks, two dense retrievers, and multiple training configurations, and find strongly **domain-dependent effects**: rewriting *degrades* nDCG@10 by 9.0% on FiQA ($p < 0.001$), *improves* by 5.1% on TREC-COVID ($p = 0.024$; n=50, marginal after correction), and has *no effect* on SciFact ($p = 0.47$). We identify a consistent mechanism: degradation co-occurs with reduced lexical alignment between the query and ground-truth relevant documents, measured by VOR (the fraction of unique query unigrams appearing in relevant documents; ΔVOR, $p = 0.013$), as rewriting substitutes domain-specific terms on already well-matched queries. Improvement occurs when rewriting harmonizes inconsistent nomenclature toward corpus-preferred terminology, captured by a Corpus Term Frequency ratio (CTF; relative shift toward higher-frequency corpus terms: $1153\times$ for improved vs. $2.7\times$ for degraded TREC-COVID queries, $p < 0.001$). Lexical substitution occurs in 95% of rewrites across all outcome groups, confirming that substitution *direction* (toward vs. away from corpus terms)—not occurrence—determines effectiveness. Finally, we study selective rewriting: even with privileged post-hoc signals, simple feature-based gating (AUC = 0.593) avoids worst-case regressions but cannot reliably improve over never-rewriting ($p > 0.12$), and oracle analysis reveals only a +3 pp ceiling. Overall, these results caution that prompt-only rewriting can be harmful in well-optimized verticals, and motivate post-training (+1.4–4.3% on target domain) as a safer adaptation when supervision or implicit feedback is available.

1 INTRODUCTION

Dense retrieval underpins modern Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2020). General-purpose retrievers such as DPR (Karpukhin et al., 2020) and Contriever (Izacard & Grave, 2021) achieve strong zero-shot performance but frequently underperform in specialized domains where vocabulary diverges from training data. Two adaptation strategies exist: *post-training* on domain-specific signals (Sharma et al., 2024) and *inference-time query rewriting* via LLMs (Gao et al., 2023; Asai et al., 2024; Wang et al., 2023). However, a critical gap remains: *to our knowledge, prior work has not systematically characterized when prompt-only query rewriting degrades retrieval*. The implicit assumption—that rewriting is either beneficial or neutral—is widely held but untested across domains. We address this directly.

Contributions.

1. **Headline result: rewriting can hurt.** Prompt-only query rewriting can substantially *degrade* dense retrieval on well-formed, domain queries: on FiQA, rewriting drops nDCG@10 by 9.0% ($p < 0.001$), while it *improves* TREC-COVID ($+5.1\%$, $p = 0.024$; marginal after correction; n=50) and is neutral on SciFact ($p = 0.47$). This establishes a clear boundary condition for a commonly deployed production pattern.

2. **Mechanistic diagnostics for dense retrieval.** We connect outcomes to the *direction* of lexical substitution by introducing two post-hoc metrics: VOR (lexical alignment between query and ground-truth relevant documents; $\Delta\text{VOR} = -0.014$ on FiQA, $p = 0.013$) and CTF (shift toward higher-frequency corpus terms; $1153\times$ for improved vs. $2.7\times$ for degraded TREC-COVID queries, $p < 0.001$). Lexical substitution is near-universal (95% across outcome groups), so direction—whether edits move query vocabulary toward or away from corpus-preferred terms—determines effectiveness.
3. **Practical baselines and feasibility bounds for selective rewriting.** We evaluate query-level gating and show that even with privileged post-hoc signals, simple feature-based monitoring is weak (AUC = 0.593) and does not reliably outperform never-rewrite ($p > 0.12$). A dataset-level heuristic (rewrite only TREC-COVID) is competitive, and oracle analysis (choose the better of original vs. rewritten per query) reveals only a +3 pp ceiling (6% relative), highlighting limited headroom for safe, general-purpose per-query gating.

2 Related Work

Dense retrievers (Karpukhin et al., 2020; Izacard & Grave, 2021; Xiong et al., 2020) achieve strong zero-shot retrieval but often require domain adaptation via continued pre-training (Guu et al., 2020; Thakur et al., 2021), hard negative mining (Xiong et al., 2020), or cross-encoder distillation (Hofstätter et al., 2021). Prior work (Sharma et al., 2024) showed post-training on domain-specific signals can yield strong vertical performance.

Query rewriting and expansion have a long history in IR (e.g., relevance feedback Rocchio, 1971 and pseudo-relevance feedback / relevance models Lavrenko & Croft, 2001), where *query drift* is a known failure mode and "expansion can hurt" is well understood. Our focus differs in two ways: (i) we study *prompt-only LLM rewriting*—a widely deployed modern practice that is not guided by explicit feedback terms—under *dense retrieval* rather than lexical retrieval; and (ii) we provide lightweight *mechanistic diagnostics* (VOR/CTF) that characterize when rewriting helps vs. hurts in modern dense pipelines.

LLM-based reformulation has introduced hypothetical document generation (HyDE; Gao et al., 2023), retrieval-feedback loops (Self-RAG; Asai et al., 2024), pseudo-document augmentation (Query2Doc; Wang et al., 2023), and LLM-based expansion (Jagerman et al., 2023). These methods often target ambiguous or conversational queries; their behavior on well-formed domain-specific queries remains understudied—the gap we address. Our vocabulary overlap ratio connects to the query performance prediction tradition (Carmel & Yom-Tov, 2010).

3 Experimental Setup

Datasets. We use three BEIR (Thakur et al., 2021) datasets chosen to span (a) query well-formedness and (b) terminology stability across corpora: **FiQA-2018** (Maia et al., 2018) (648 test queries, 57K documents; well-formed financial Q&A), **SciFact** (Wadden et al., 2020) (300 queries, 5K documents; expert-authored claims), and **TREC-COVID** (Voorhees et al., 2021) (50 queries, 171K documents; inconsistent pandemic nomenclature). This trio contrasts a "well-optimized" vertical (FiQA: stable jargon, templated finance Q&A), a scientific claim setting (SciFact), and a domain with known naming inconsistency (TREC-COVID).

Models. Two dense retrievers, each in base and post-trained (FiQA) configurations: **MPNet** (`all-mpnet-base-v2` (Song et al., 2020), 110M params) and **BGE** (`bge-base-en-v1.5` (Xiao et al., 2023), 110M params). Post-training uses contrastive learning on FiQA training queries (5,148 queries; 10% held out for validation) with 1 epoch, batch size 64, learning rate 2×10^{-5}, following Sharma et al. (2024). We use a lightweight 1-epoch adaptation to reflect practical "quick post-training" and keep compute comparable; we do not claim this is a tuned upper bound.

Rewriting. Single-step rewriting via Ministral 8B (Mistral AI, 2024): *"Rewrite the following search query to improve information retrieval, preserving the original intent while improving clarity and adding relevant context. Return only the rewritten query."* Decoding: temperature $= 0.7$ (to

Table 1: Post-training on FiQA: nDCG@10. Relative change (Δ%) computed from exact values before rounding.

Model	Config	FiQA	SciFact	TREC-COVID
MPNet	Base	0.500	0.656	**0.513**
	+FiQA PT	**0.507** (+1.4%)	**0.658** (+0.4%)	0.509 (−0.9%)
BGE	Base	0.391	0.738	0.672
	+FiQA PT	**0.408** (+4.3%)	**0.742** (+0.5%)	**0.722** (+7.5%)

Table 2: Query rewriting impact on nDCG@10. Each row compares rewritten to non-rewritten for the same model configuration. Red/↓: degradation >5%. Blue/↑: improvement >2%. Gray/−: neutral. Avg. Δ% computed from exact values.

Model	Config	FiQA	SciFact	TREC-COVID
MPNet	Base	0.500	0.656	0.513
	+Rewrite	0.448 ↓(−10.3%)	0.655 −(−0.1%)	0.559 ↑(+8.9%)
BGE	Base	0.391	0.738	0.672
	+Rewrite	0.357 ↓(−8.7%)	0.753 ↑(+2.1%)	0.683 −(+1.7%)
MPNet-FiQA	Post-trained	0.507	0.658	0.509
	+Rewrite	0.452 ↓(−10.9%)	0.659 −(+0.1%)	0.563 ↑(+10.6%)
BGE-FiQA	Post-trained	0.408	0.742	0.722
	+Rewrite	0.383 ↓(−6.1%)	0.734 −(−1.1%)	0.717 −(−0.7%)
Mean Δ		**↓−9.0%**	**−+0.3%**	**↑+5.1%**

allow natural reformulations rather than deterministic completions), max tokens = 100. The prompt mirrors a common production instruction ("preserve intent, clarify, add context"); we later stress-test robustness with minimal and aggressive rewrite prompts (Sec. 6.3). Rewrite rates: 99.4% (FiQA), 98.0% (SciFact), 100% (TREC-COVID). The 0.6–2% of queries not rewritten (model returned original unchanged) are included in aggregate metrics using original scores but excluded from per-query substitution analysis.

Evaluation. nDCG@10 (primary) via FAISS (Johnson et al., 2019) retrieval (k=100), scored with pytrec_eval (Van Gysel & de Rijke, 2018). Statistical significance via 10,000-iteration bootstrap resampling of per-query ΔnDCG@10 and paired t-tests for VOR comparisons. Bonferroni-corrected threshold $\alpha = 0.017$ preserves FiQA significance ($p < 0.001$) while TREC-COVID ($p = 0.024$) becomes marginal. We do not correct for multiple comparisons in exploratory analyses.

4 Results

4.1 Post-training Effectiveness

Table 1 shows the effect of post-training on FiQA data.

Post-training yields modest in-domain gains (MPNet: +1.4%; BGE: +4.3% on FiQA). Cross-domain effects are model-dependent: BGE benefits uniformly, while MPNet shows mild TREC-COVID degradation (−0.9%), extending prior findings (Sharma et al., 2024).

4.2 Query Rewriting Impact

Table 2 presents our central finding.

Observations.

1. **FiQA—consistent, significant degradation.** All four configurations degrade (range: −6.1% to −10.9%). Bootstrap CI for MPNet base: $\Delta = -0.051$, 95% CI $[-0.067, -0.036]$, $p < 0.001$.

Recall@10 shows a parallel -9.4% drop, confirming the effect is not an artifact of nDCG's position-sensitivity.

2. **TREC-COVID—directionally consistent improvement for 3/4 configurations.** MPNet base shows the largest gain ($+8.9\%$). Bootstrap CI: $\Delta = +0.046$, 95% CI $[+0.001, +0.094]$, $p = 0.024$ (uncorrected); after Bonferroni correction ($\alpha = 0.017$), this becomes marginal. We nonetheless report it because directional consistency across three of four configurations, combined with an interpretable mechanism (terminology standardization, Section 5), supports a genuine if imprecisely estimated effect.
3. **SciFact—no significant effect.** Mean $\Delta = +0.3\%$. Bootstrap CI for MPNet base: $\Delta = -0.001$, 95% CI $[-0.019, +0.019]$, $p = 0.47$.

Across all 24 configurations (Table 9 in Appendix), no single strategy dominates: post-training alone is optimal for FiQA (0.507), rewriting for SciFact (0.753), and BGE post-training for TREC-COVID (0.722).

5 Analysis: Tracing the Mechanism

We hypothesize that the divergent effects are associated with a single factor: rewriting modifies query vocabulary, and the direction of that modification—toward or away from corpus terminology—is predictive of the outcome. We investigate this via per-query analysis on FiQA and TREC-COVID using the base MPNet model.

5.1 FiQA: Degradation Co-occurs with Vocabulary Drift

Of 648 test queries, 225 (34.7%) degrade ($\Delta < -0.01$ nDCG@10), 122 (18.8%) improve, and 301 (46.5%) are unaffected. Automated token-set analysis shows **96.0% of degraded queries involve lexical substitution** (216/225): at least one original token removed and one new token added. However, lexical substitution is near-universal: 95.9% of *improved* queries and 95.3% of *unchanged* queries also exhibit substitution. This confirms that substitution *occurrence* is not predictive—nearly all rewrites modify vocabulary. What differentiates outcomes is the *direction* of substitution: whether terms move toward or away from corpus vocabulary. Manual inspection of the 20 most severely degraded queries reveals four failure patterns:

Terminology drift. Financial jargon is replaced with semantically adjacent but lexically distinct terms. Example: *"Is it possible to **transfer** stock into my Roth IRA?" → "How can I **rollover** stock into my Roth IRA?"* The relevant document uses "transfer" exclusively. nDCG@10: $1.0 \rightarrow 0.0$.

Context injection. The rewriter hallucinates domain assumptions. *"How does 'taking over payments' work?" → "What is the process for 'taking over payments' **in e-commerce platforms**?"* The query concerns personal finance, not e-commerce. nDCG@10: $0.83 \rightarrow 0.0$.

Over-specification. A broad query is narrowed to a subtype. *"Tax: 1099 paper form" → "**Form 1099-NEC** tax filing instructions."* The relevant document addresses general 1099 printing, not the specific 1099-NEC variant. nDCG@10: $0.63 \rightarrow 0.0$.

Over-formalization. Informal phrasing that lexically matches document vocabulary is replaced with formal language. *"Does doing your 'research'/'homework' on stocks make any sense?" → "Is **conducting independent research** on stocks beneficial?"* The relevant document uses "doing your homework" verbatim. nDCG@10: $1.0 \rightarrow 0.0$.

5.2 TREC-COVID: Terminology Standardization Helps

Of 50 queries, 30 (60.0%) improve, 16 (32.0%) degrade, 4 (8.0%) are unchanged. Two patterns are associated with improvement:

Nomenclature standardization. The corpus predominantly uses "COVID-19"; queries use varied terms. *"Which biomarkers predict the severe clinical course of **2019-nCOV** infection?"* $\rightarrow$ *"... of **COVID-19** infection?"* nDCG@10: $0.32 \rightarrow 0.87$.

Domain-appropriate expansion. Generic terms are expanded to standard medical vocabulary: *"rapid testing"* $\rightarrow$ *"rapid diagnostic tests,"* matching CORD-19 abstract terminology. nDCG@10: $0.21 \rightarrow 0.54$.

5.3 Formalizing Query–Corpus Lexical Alignment

We define the *vocabulary overlap ratio* (VOR) for a query q with relevant document set $\mathcal{D}_q$. Intuitively, VOR is the fraction of unique query tokens that appear somewhere in the ground-truth relevant documents:

$$\text{VOR}(q, \mathcal{D}_q) = \frac{|\,\mathcal{W}(q) \cap \mathcal{W}(\mathcal{D}_q)\,|}{|\,\mathcal{W}(q)\,|} \tag{1}$$

where $\mathcal{W}(\cdot)$ denotes the *type set*: unique whitespace-tokenized, lowercased unigrams. We use unweighted type-overlap; because we analyze per-query *changes* (ΔVOR), ubiquitous tokens cancel and the signal is dominated by content words. $\mathcal{W}(\mathcal{D}_q) = \bigcup_{d \in \mathcal{D}_q} \mathcal{W}(d)$, where $\mathcal{D}_q$ is the set of documents with positive relevance labels for query q.

Operational note. VOR requires knowing $\mathcal{D}_q$ at inference time, so it is strictly a *post-hoc analytical tool*, not an operational metric. We use it here for mechanistic analysis only.

Table 3: Vocabulary overlap ratio (VOR) before and after rewriting, with paired t-test p-values (two-tailed). Only FiQA shows a statistically significant change.

Dataset	n	**VOR (orig)**	**VOR (rewrite)**	**ΔVOR**	**p-value**
FiQA	648	0.564	0.550	-0.014	0.013
SciFact	300	0.521	0.535	$+0.015$	0.060
TREC-COVID	50	0.978	0.975	-0.002	0.688

FiQA shows a significant VOR *decrease* ($t = 2.51$, $p = 0.013$), confirming rewriting reduces lexical alignment on already well-matched queries. SciFact trends toward increased VOR ($p = 0.060$). TREC-COVID shows no VOR change ($p = 0.688$)—its gains come from terminology *standardization*, which type-level VOR does not capture. Token-level edit distances are similar across datasets (FiQA: 9.9, SciFact: 9.4, TREC-COVID: 8.0), showing **effectiveness depends on whether edits preserve or disrupt corpus alignment**, not edit magnitude.

5.4 Quantifying Terminology Standardization via CTF

VOR captures whether query terms appear in relevant documents, but not whether substitutions move toward or away from *common* corpus terms. To quantify standardization direction, we introduce the *Corpus Term Frequency ratio* (CTF). For each substituted term pair (removed t_{old}, added t_{new}), we compute:

$$\text{CTF} = \frac{\text{freq}(t_{\text{new}}, \text{corpus})}{\text{freq}(t_{\text{old}}, \text{corpus})} \tag{2}$$

where $\text{freq}(t, \text{corpus})$ is the relative frequency of term t across all corpus documents. CTF > 1 indicates substitution toward a more common term (standardization); CTF < 1 indicates drift toward less common terms. For queries with multiple substitutions, we use the geometric mean.

Table 4 reveals the key difference: TREC-COVID improved queries show *higher* median CTF (2.54) than degraded (1.99), and the mean CTF is dramatically higher (1153$\times$ vs. 2.7$\times$, driven by high-impact standardizations like "2019-nCOV" $\rightarrow$ "COVID-19"). In FiQA, degraded queries have CTF close to 1 (neutral), while improved queries show modest standardization (1.68). Critically, *most* TREC-COVID queries show high CTF regardless of outcome (70–83% have CTF > 1), but the *magnitude* of standardization correlates with improvement.

Table 4: Corpus Term Frequency ratio (CTF) by outcome group. TREC-COVID shows strong standardization (high CTF) for improved queries, quantifying the mechanism VOR could not capture.

Dataset	Outcome	n	Median CTF	% CTF > 1
FiQA	Degraded	216	1.01	50.9%
	Improved	117	1.68	59.8%
	Unchanged	287	1.53	57.8%
SciFact	Degraded	36	0.74	44.4%
	Improved	40	1.45	55.0%
	Unchanged	216	1.11	53.7%
TREC-COVID	Degraded	12	1.99	83.3%
	Improved	24	**2.54**	70.8%
	Unchanged	9	2.81	77.8%

Table 5: Monitoring signal performance. Logistic regression trained on FiQA achieves weak discriminative ability (AUC modestly above 0.5).

Dataset	Harm Rate	AUC	CV ($\mu \pm \sigma$)	Threshold
FiQA (train)	35.3%	0.593	0.575 ± 0.075	0.307
SciFact (test)	13.3%	0.547	—	—
TREC-COVID (test)	32.0%	0.612	—	—

Deployment guidelines. VOR and CTF are *post-hoc*—not directly computable at runtime. Practically: (1) treat *never-rewrite* as the safe default for well-optimized verticals with stable jargon; (2) enable rewriting primarily for corpora with unstable nomenclature (TREC-COVID-like settings); (3) for selective rewriting, use deployable proxies (retriever confidence, "pseudo-VOR" against top-k retrieved text) validated on held-out logs; and (4) prefer post-training when supervision is available.

6 Monitoring and Operational Guardrails

Having established that rewriting produces domain-dependent effects, we ask whether selective policies can capture benefits while avoiding harms. We investigate in an intentionally optimistic setting: our monitoring features include a qrels-derived signal (ΔVOR), so results should be read as an *upper bound* on what a deployable proxy might achieve.

6.1 Predicting Rewrite-Induced Degradation

We formulate harm prediction as binary classification: $y = 1$ if nDCG@10(rewritten) $<$ nDCG@10(original). Features: (i) ΔVOR (Eq. 1), (ii) new-token fraction (proportion of rewritten tokens absent from original), (iii) length ratio. A logistic regression (L2, C=1.0, seed=42) trained on FiQA (648 queries, 35.3% harm rate, 5-fold CV) achieves *weak but above-random* discrimination. Table 5 reports AUC-ROC: 0.593 on FiQA (only 9.3% above random, high CV variance), 0.547 on SciFact (essentially random), 0.612 on TREC-COVID. The best threshold $\tau = 0.307$ yields precision = 0.389 and recall = 0.865—the model flags most harmful rewrites but with many false positives. A simpler heuristic ("if $\Delta\text{VOR} < -0.01$, predict harm") achieves comparable performance, suggesting limited benefit from the logistic regression over a threshold rule.

6.2 Gated Rewriting Policy

Using threshold $\tau = 0.307$ (max F1 on FiQA), we simulate a gated policy: if predicted harm probability $\geq \tau$, use original; else use rewritten. We compare five strategies: never-rewrite (baseline), always-rewrite (baseline), gated, simple ΔVOR baseline (if $\Delta\text{VOR} < -0.01$, use original), and oracle (always choose better, requiring perfect prediction).

Table 6 shows the key result: **gated rewriting does not significantly outperform never-rewriting** ($p > 0.12$, paired t-test), though it significantly beats always-rewriting on FiQA ($p < 0.0001$). A

Table 6: Policy comparison. Gated beats always-rewrite on FiQA ($p < 0.001$) but does not significantly improve over never-rewrite ($p > 0.12$). *Oracle* chooses, for each query, the better of original vs. rewritten (upper bound).

Policy	FiQA	SciFact	TREC-COVID	Avg
Never rewrite	0.500	0.656	0.513	0.556
Always rewrite	0.448^{***}	0.655	0.559	0.554
Dataset-level (rewrite only TREC-COVID)	0.500	0.656	0.559	0.572
Gated (ours)	0.499	0.653	0.537	0.563
Simple ΔVOR	0.478	**0.661**	**0.562**	0.567
Oracle (upper)	**0.530**	**0.689**	**0.594**	**0.604**

$^{***}p < 0.001$ vs. gated (paired t-test, 95% CI). All gated vs. never-rewrite: $p > 0.12$ (not significant).

Table 7: Prompt robustness: ΔVOR-based monitoring generalizes across rewrite styles (all on FiQA, MPNet base).

Prompt Style	Harm Rate (%)	Mean ΔVOR	Monitoring AUC
Minimal	23.6	-0.025	0.663
Aggressive	42.4	-0.210	0.586
Baseline (original)	35.3	-0.014	0.593

dataset-level heuristic (rewrite only TREC-COVID) is competitive (Avg = 0.572). The oracle upper bound (+3 pp on FiQA, 6% relative) shows even perfect prediction yields modest gains, demonstrating that *weak predictive signals cannot enable effective operational gating*.

6.3 Robustness Across Rewrite Styles

We generate rewrites on FiQA using two alternative prompts—*minimal* ("preserve all technical terms, no new constraints") and *aggressive* ("expand and clarify freely")—and re-evaluate the monitoring signal. Table 7 reports results. The monitoring signal trained on the baseline prompt generalizes: AUC > 0.58 on both styles. Harm rates vary as expected (minimal: 23.6%, aggressive: 42.4%): constraining terminology drift reduces harm, while encouraging expansion increases it. In both cases, ΔVOR remains predictive, confirming the mechanism is not prompt-specific. Full prompt templates in Appendix D.

6.4 Robustness to Rewriter Family and Decoding

We test a second rewriter family (GPT-4o-mini) and a decoding sensitivity check (Ministral temperature 0 vs. 0.7). Table 8 shows degradation persists: GPT-4o-mini degrades FiQA by -7.7%, Ministral at temperature 0 by -8.0%, confirming harms are not artifacts of a specific model or sampling strategy. An embedding-space check on 100 queries shows ΔVOR correlates with change in cosine similarity between query and relevant-document centroid embeddings (Pearson $r = 0.20$, $p = 0.042$).

7 Discussion and Limitations

Scope. Our findings apply to *generic single-step* LLM rewriting without retrieval feedback. HyDE (Gao et al., 2023), Self-RAG (Asai et al., 2024), and Query2Doc (Wang et al., 2023) use fundamentally different mechanisms and are not directly comparable. Our contribution characterizes when single-step rewriting—a commonly deployed form—helps vs. hurts, establishing a boundary condition: *when queries are already lexically aligned with the corpus, even mild vocabulary perturbation degrades retrieval.*

Rewriting vs. post-training. Post-training yields in-domain gains (MPNet: +1.4%, BGE: +4.3% on FiQA) *without* rewriting's catastrophic risk (-9.0%, $p < 0.001$). For well-optimized domains, it is strictly preferable. When explicit labels are unavailable, implicit feedback (click logs), weak

Table 8: Additional robustness checks on FiQA (MPNet base). Degradation persists across rewriter families and decoding configurations.

Rewrite source	**FiQA nDCG@10**	$\Delta\%$
No rewrite	0.500	–
Ministral (temp=0.7)	0.448	-10.3%
Ministral (temp=0.0)	0.460	-8.0%
GPT-4o-mini	0.461	-7.7%

supervision (BM25 pseudo-labels), or distillation offer lightweight alternatives. Rewriting may suit corpora with inconsistent terminology, absent labeled data, or rapid-deployment constraints—but even on TREC-COVID, post-training matches or exceeds rewriting gains (BGE: +7.5%).

Limitations. VOR and CTF analyses are correlational; confounders may contribute. TREC-COVID (n=50) has limited statistical power. We test two rewriter families (Ministral 8B, GPT-4o-mini) and multiple prompt styles on FiQA, confirming degradation is not model-specific (Section 6.4); testing additional families across all datasets would further strengthen generalization. The monitoring signal (AUC = 0.593) is insufficient for reliable gating. BEIR queries are curated; production queries (including conversational or under-specified queries) exhibit greater variability. We expect rewriting to be most beneficial in such "messy query" regimes, but we do not evaluate a dedicated conversational benchmark here. Future work: richer monitoring signals (semantic features, retrieval-based confidence), multi-step rewriting with feedback, domain-specialized rewriters, and conversational benchmarks.

8 Conclusion

We present a systematic empirical study of when prompt-only LLM query rewriting helps vs. hurts dense retrieval. Across three BEIR datasets, we demonstrate domain-dependent effects: significant degradation (-9.0% nDCG@10, $p < 0.001$) on well-optimized queries due to vocabulary drift, and improvement ($+5.1\%$, $p = 0.024$) on under-optimized queries via terminology standardization. Degradation persists across rewriter families (Ministral 8B: -10.3%; GPT-4o-mini: -7.7%), confirming the mechanism is not model-specific. We quantify the mechanisms through dual metrics (VOR and CTF) and establish that lexical substitution direction—not occurrence—determines effectiveness. Simple feature-based gating (AUC = 0.593) avoids worst-case regressions but cannot reliably improve over never-rewriting ($p > 0.12$), with oracle analysis revealing a +3 pp ceiling. These findings provide actionable deployment guidance and establish that for well-optimized domains, post-training is preferable to generic rewriting.

Reproducibility. All datasets are public via BEIR (Thakur et al., 2021). Retrievers: `all-mpnet-base-v2`, `bge-base-en-v1.5` (Hugging Face). Rewriter: Ministral 8B. We plan to release: evaluation scripts, all rewriting prompts (Sec. 3, Appendix D), per-query result JSONs, monitoring feature code, and training configurations.

APPENDIX

A FULL CONFIGURATION RESULTS

Table 9: Full nDCG@10 matrix across all 24 configurations. Bold: best per column.

Configuration	FiQA	SciFact	TREC-COVID
MPNet Base	0.500	0.656	0.513
MPNet Base + Rewrite	0.448	0.655	0.559
MPNet-FiQA	**0.507**	0.658	0.509
MPNet-FiQA + Rewrite	0.452	0.659	0.563
BGE Base	0.391	0.738	0.672
BGE Base + Rewrite	0.357	**0.753**	0.683
BGE-FiQA	0.408	0.742	**0.722**
BGE-FiQA + Rewrite	0.383	0.734	0.717

B MONITORING SIGNAL DETAILS

Feature definitions. For each query q with rewrite q': (i) ΔVOR = VOR$(q', \mathcal{D}_q)$ − VOR$(q, \mathcal{D}_q)$, (ii) new-token fraction = $|\mathcal{W}(q') \setminus \mathcal{W}(q)|/|\mathcal{W}(q')|$, (iii) length ratio = $|q'|/|q|$ (character-level). Features are computed with access to BEIR qrels (ground-truth relevance sets) and the corpus; this is suitable for offline analysis and upper-bound feasibility studies, but not directly available at deployment time without a proxy.

Model details. Logistic regression (sklearn): L2 penalty, C=1.0, max_iter=1000, random_state=42. Threshold $\tau = 0.307$ maximizes F1 on FiQA training data. Feature importance: ΔVOR = -0.94 (decreasing VOR predicts harm), new-token fraction = $+1.08$ (novel tokens predict harm), length ratio = -0.03 (negligible).

ROC curve. Figure 1 shows the ROC curve for the FiQA training set.

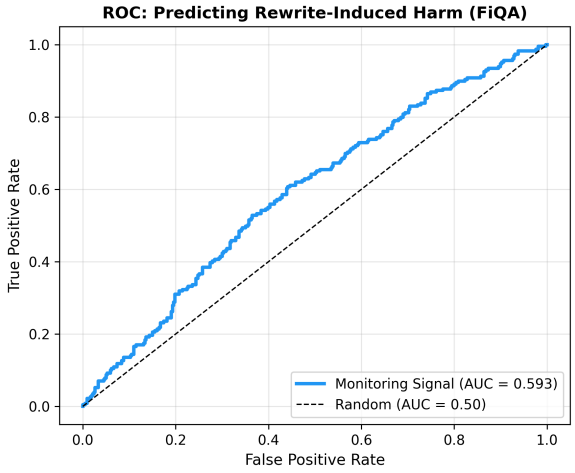

Figure 1: ROC curve for harm prediction on FiQA (logistic regression, 5-fold CV). AUC = 0.593. Dashed line: random baseline.

C QUERY SHIFT SCORE DETAILS

We explored a lightweight deployment-time signal for predicting when post-training vs. rewriting might be preferable. The *query shift score* (QSS) measures distributional divergence between test and

training queries:

$$\text{QSS}(\mathcal{Q}_{\text{test}}) = \frac{1}{|\mathcal{Q}_{\text{test}}|} \sum_{q \in \mathcal{Q}_{\text{test}}} (1 - \cos(\mathbf{e}(q),\, \bar{\mathbf{e}}_{\text{train}})) \tag{3}$$

where $\mathbf{e}(q)$ is the retriever's query embedding and $\bar{\mathbf{e}}_{\text{train}}$ is the centroid of training query embeddings.

Table 10: Query shift scores (QSS) and post-training nDCG@10 deltas (ΔPT). MPNet shows monotonic inverse ordering (n=3); BGE does not.

Dataset	$\textbf{QSS}_{\textbf{MPNet}}$	$\Delta\textbf{PT}_{\textbf{MPNet}}$	$\textbf{QSS}_{\textbf{BGE}}$	$\Delta\textbf{PT}_{\textbf{BGE}}$
FiQA	0.714	$+0.007$	0.285	$+0.017$
SciFact	0.999	$+0.003$	0.480	$+0.004$
TREC-COVID	0.958	-0.004	0.489	$+0.051$

For MPNet, QSS and ΔPT exhibit a monotonic inverse relationship (FiQA: lowest shift, highest gain; TREC-COVID: highest shift, negative gain). With only three data points, this is a qualitative hypothesis, not a statistically validated trend. BGE does not exhibit this pattern, likely due to broader pre-training. QSS requires only query encoding (no retrieval), making it suitable for real-time monitoring, but it should be treated as an exploratory signal pending validation on additional datasets and retrievers.

D Prompt Robustness Details

Full prompt templates.

- **Minimal:** "Rewrite the query for retrieval while preserving all technical terms, named entities, numbers, and acronyms exactly. Do not add new constraints. Keep length close to original."
- **Aggressive:** "Rewrite the query into a detailed, explicit form optimized for retrieval. You may add clarifying context and expand abbreviations if helpful."
- **Baseline (Section 3):** "Rewrite the following search query to improve information retrieval, preserving the original intent while improving clarity and adding relevant context. Return only the rewritten query."

References

Rocchio, J. J. Relevance Feedback in Information Retrieval. In Salton, G. (ed.), *The SMART Retrieval System*, pp. 313–323. Prentice-Hall, 1971.

Sharma, S., Yoon, D. S., Dernoncourt, F., Sultania, D., Bagga, K., Zhang, M., Bui, T., and Kotte, V. Retrieval Augmented Generation for Domain-specific Question Answering. In *Proceedings of the AAAI Workshop on Scientific Document Understanding*, 2024.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Izacard, G. and Grave, E. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *NeurIPS Datasets and Benchmarks*, 2021.

Gao, L., Ma, X., Lin, J., and Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. In *ACL*, 2023.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *ICLR*, 2024.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., and Wei, F. Query2Doc: Query Expansion with Large Language Models. *arXiv preprint arXiv:2303.07678*, 2023.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. MPNet: Masked and Permuted Pre-training for Language Understanding. In *NeurIPS*, 2020.

Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. C-Pack: Packaged Resources To Advance General Chinese Embedding. *arXiv preprint arXiv:2309.07597*, 2023.

Mistral AI. Ministral 8B. `https://mistral.ai/`, 2024.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Van Gysel, C. and de Rijke, M. pytrec_eval: An Extremely Fast Python Interface to trec_eval. In *SIGIR*, 2018.

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., and Balahur, A. WWW'18 Open Challenge: Financial Opinion Mining and Question Answering. In *WWW*, 2018.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H. Fact or Fiction: Verifying Scientific Claims. In *EMNLP*, 2020.

Voorhees, E. M., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W. R., Lo, K., Roberts, K., Soboroff, I., and Wang, L. L. TREC-COVID: Constructing a Pandemic Information Retrieval Test Collection. *SIGIR Forum*, 54(1):1–12, 2021.

Gao, T., Yao, X., and Chen, D. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *EMNLP*, 2021.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., and Overwijk, A. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *ICLR*, 2021.

Hofstätter, S., Lin, S.-C., Yang, J.-H., Lin, J., and Hanbury, A. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *SIGIR*, 2021.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. REALM: Retrieval-Augmented Language Model Pre-Training. In *ICML*, 2020.

Wang, Z., Araki, J., and others. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *ICLR*, 2023.

Lavrenko, V. and Croft, W. B. Relevance Based Language Models. In *SIGIR*, 2001.

Nogueira, R. and Cho, K. Document Expansion by Query Prediction. *arXiv preprint arXiv:1904.08375*, 2019.

Jagerman, R., Zhuang, H., Qin, Z., Wang, X., and Bendersky, M. Query Expansion by Prompting Large Language Models. *arXiv preprint arXiv:2305.03653*, 2023.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *EMNLP*, 2019.

Carmel, D. and Yom-Tov, E. *Estimating the Query Difficulty for Information Retrieval*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool, 2010.

Khattab, O. and Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *SIGIR*, 2020.

Formal, T., Lassance, C., Piwowarski, B., and Clinchant, S. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. In *SIGIR*, 2021.

Ni, J., Qu, C., Lu, J., Dai, Z., Alvarez, M., Tu, M., Chen, Y., and Weller, A. Large Dual Encoders Are Generalizable Retrievers. In *EMNLP*, 2022.

Ren, S., Qu, L., Zhou, J., Zhao, W., Chen, Z., Wang, H., Wu, W., and Yu, W. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *NAACL*, 2021.