The Wisdom of Many Queries: Complexity-Diversity Principle for Dense Retriever Training

Anonymous ACL submission

Abstract

Prior work reports conflicting results on query diversity in synthetic data generation for dense retrieval. We identify this conflict and design Q-D metrics to quantify diversity's impact, making the problem measurable. Through experiments on 4 benchmark types (31 datasets), we find query diversity especially benefits multi-hop retrieval. Deep analysis on multi-hop data reveals that diversity benefit correlates strongly with query complexity ($r \geq 0.95$, $p < 0.05$ in 12/14 conditions), measured by content words (CW). We formalize this as the **Complexity-Diversity Principle (CDP)**: *query complexity determines optimal diversity*. CDP provides actionable thresholds (CW>10: use diversity; CW<7: avoid it). Guided by CDP, we propose zero-shot multi-query synthesis for multi-hop tasks, achieving state-of-the-art performance.

1 Introduction

Task: Synthetic Query Generation for Dense Retrieval. Dense retrieval models encode queries and documents into a shared embedding space for efficient similarity search (Karpukhin et al., 2020; Xiong et al., 2021). Training these models requires large-scale query-document pairs, which are expensive to annotate manually. *Synthetic query generation* addresses this by using LLMs to generate training queries from documents (Bonifacio et al., 2022; Dai et al., 2023). Given a document corpus $\mathcal{D}$, the goal is to generate queries $\{q_i\}$ such that training on (q_i, d_i) pairs yields a retriever that generalizes well to unseen evaluation tasks.

Prior Work: Focus on Query Quality. Existing methods predominantly focus on generating *high-quality* queries, i.e., those resembling human-written queries in style and content. InPars (Bonifacio et al., 2022) uses few-shot prompting with MS MARCO examples. Promptagator (Dai et al., 2023) incorporates target-domain examples. SAP (Thakur et al., 2024) proposes a summarize-then-ask pipeline. These methods share a common assumption: higher query quality leads to better retriever performance.

Query: What does Ivan promise to do when he turns thirty?

Q/Doc=1 ✗ *"I do love you, Ivan. Dmitri says of you—Ivan is a tomb!"*
↪ Matches character name only

Q/Doc=3 ✓ *"Listen, Alyosha," Ivan began, "if I am really able to care for the sticky little leaves..."*
↪ Ivan's actual promise about living until thirty

Figure 1: Diverse queries act as regularization (Contriever case study). With single-query training (Q/Doc=1), the retriever overfits to surface features like character names, retrieving passages that merely *mention* "Ivan." With diverse training queries (Q/Doc=3), the model learns fine-grained semantic matching and correctly retrieves passages about Ivan's *actual promise*. Example from NovelHopQA (Gupta et al., 2025).

The Unexplored Role of Query Diversity. However, the role of query *diversity*, i.e., generating multiple varied queries per document, remains largely unexplored. While some methods incidentally generate multiple queries (Dai et al., 2023; Lin et al., 2023), no prior work has systematically studied *when* and *why* diversity helps. Anecdotal evidence is conflicting: some report diversity helps out-of-domain (OOD) generalization; others find it harmful for in-domain tasks. This gap motivates a fundamental question: *when does query diversity help, and when does it hurt?*

Our Hypothesis. We hypothesize that the conflicting results stem from ignoring a key factor: **query complexity**. Our intuition is that complex queries, i.e., those expressing rich information needs with multiple entities, relationships, or conditions, can be formulated in many semantically distinct ways. Training on diverse formulations helps the retriever learn this rich semantic space (Figure 1). In contrast, simple queries have limited paraphrase possibilities; excessive diversity may introduce noise by straying from the core information need. If this hypothesis holds, diversity benefit should correlate with query complexity.

Contributions.

- **Problem Formulation**: We identify conflicting diversity needs across tasks and design Q-D metrics to quantify diversity's impact on retriever training, making the problem measurable.
- **Broad Validation**: Through experiments on 4 benchmark types (31 datasets), we find that query diversity is especially beneficial for multi-hop retrieval tasks.
- **Deep Analysis**: On multi-hop data, we discover the Complexity-Diversity Principle (CDP), showing that query complexity determines optimal diversity ($r \geq 0.95$, $p < 0.05$), and propose zero-shot multi-query synthesis guided by CDP.

2 Related Work

Query Synthesis for Dense Retrieval The scarcity of labeled data has motivated research on synthetic query generation for training dense retrievers. InPars (Bonifacio et al., 2022) pioneered the use of LLMs for query synthesis, demonstrating that few-shot prompting with GPT-3 can generate effective training queries. Promptagator (Dai et al., 2023) extended this idea by incorporating target task examples and introduced round-trip consistency filtering to improve query quality. SWIM-IR (Thakur et al., 2024) proposed a summarize-then-ask pipeline (SAP) to enhance query quality across multiple languages. DRAGON (Lin et al., 2023) combined sentence-level cropping with diverse supervision sources. Promptodile (Gwon et al., 2025) explored using smaller open-source LLMs for Promptagator-style training, finding that filtering may become less important as LLMs improve. While these methods may incidentally produce multiple queries, **none systematically investigates when query diversity helps or hurts**, which is the central question we address in this work.

Data Augmentation for Retrieval Beyond query synthesis, various data augmentation strategies have been explored for training retrievers. Contriever (Izacard et al., 2022) and RetroMAE (Xiao et al., 2022) leverage unsupervised pretraining objectives. DPR (Karpukhin et al., 2020) introduced the dual-encoder paradigm with in-batch and BM25-mined hard negatives, while ANCE (Xiong et al., 2021) further improved hard negative sampling via approximate nearest neighbor search. Our work is orthogonal to these approaches and can be combined with them.

Out-of-Domain Generalization Improving out-of-domain (OOD) generalization is a critical challenge for dense retrievers. BEIR (Thakur et al., 2021) established a benchmark for evaluating zero-shot transfer capabilities. While prior work has explored pretraining objectives (Izacard et al., 2022; Xiao et al., 2022) and training data scaling to improve generalization, the specific role of *query diversity* remains unstudied. Our work fills this gap by providing the first systematic analysis of how query diversity during training affects OOD performance.

3 Methodology

Our primary contribution is the Complexity-Diversity Principle (CDP), not the query generation method itself. To test our hypothesis, we adopt a **zero-shot multi-query synthesis** approach that explicitly controls diversity through prompt instructions. Unlike few-shot methods that produce homogeneous queries due to pattern copying, our zero-shot design: (1) varies diversity by simply changing prompts, (2) generates M queries per document in a single LLM call, and (3) isolates diversity effects from prompt engineering artifacts. Figure 2 provides an overview (pseudocode in Appendix A).

3.1 Step 1: Zero-shot Multi-Query Synthesis

The core of our framework is generating multiple diverse queries per document using a single zero-shot LLM call. Unlike few-shot methods that produce homogeneous queries due to pattern copying (CE>0.5, see Table 1 and Appendix K), we explicitly instruct the LLM to generate M semantically diverse queries with temperature=0.

The prompt specifies multiple query formats (Figure 3): factual (What), procedural (How), causal (Why), conditional (When/If), keyword

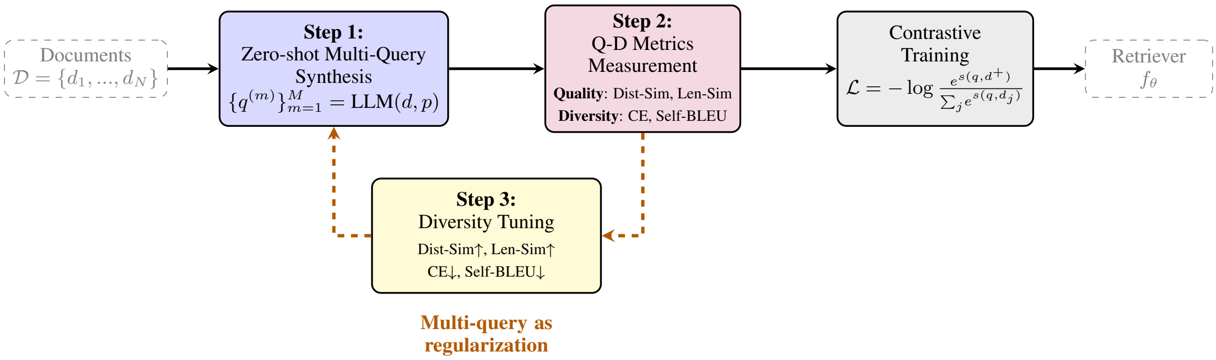

Figure 2: Experimental pipeline for testing the Complexity-Diversity Principle. Given a document corpus, we (1) generate M diverse queries per document using zero-shot prompting, (2) measure query quality and diversity using Q-D metrics, and (3) tune diversity level based on target task (dashed arrow shows iteration on sample data). The retriever is then trained with standard contrastive loss. This pipeline enables controlled experiments varying diversity while holding other factors constant.

queries, statement/claim format, and comparison questions. This instruction-based approach achieves: (1) **True semantic diversity**, where queries differ in both form and content, not just surface wording; (2) **Reproducibility**, as temperature=0 ensures deterministic outputs; (3) **Cost efficiency**, since generating M queries in one call dramatically reduces API costs.

Diverse Prompt for Multi-hop Tasks

Your task is to generate {M} independent queries based on the document(s).

You MUST generate queries in these specific formats:
– **What**... questions (factual)
– **How**... questions (procedural)
– **Why**... questions (causal)
– **When/If**... questions (conditional)
– **Keyword** queries (2-5 words, no question mark)
– **Statement/claim** format (e.g., "X is used for Y")
– **Comparison** questions

Each query must target different information from the document.

Figure 3: Diverse prompting strategy for multi-hop tasks. The prompt enforces varied query formats (factual, procedural, causal, conditional, keyword, statement, comparison) targeting different information from the document. M denotes Q/Doc.

3.2 Step 2: Q-D Metrics Measurement

We propose Quality-Diversity (Q-D) metrics to quantify synthetic queries. For *quality*: (1) **Dist-Sim**, the cosine similarity between synthetic and human query embeddings, indicating semantic alignment; (2) **Len-Sim**, the ratio of query lengths, capturing stylistic similarity. For *diversity*: (1) **CE** (Cross-Encoder Paraphrase Ratio), the proportion of query pairs classified as paraphrases, where lower values indicate more semantic diversity; (2) **Self-BLEU**, the average BLEU score between query pairs, where lower values indicate more lexical diversity.

3.3 Step 3: Diversity Tuning

Different tasks benefit from different diversity levels. We provide two prompt variants: (1) **Paraphrase mode** generates M semantically equivalent queries (same question, different wording; CE≈0.50), better for in-domain tasks; (2) **Diverse mode** generates M semantically distinct queries (different formats; CE≈0.04), better for reasoning-intensive OOD tasks. Full templates in Appendix L; training details in Appendix C.

4 Experimental Setup

Training Data. We use an 80k subset of MS MARCO QA v1.1 (Nguyen et al., 2016) as our source corpus. For each document, we generate synthetic queries using different methods and train dense retrievers.

Backbone LLM. We use GPT-4o-mini (OpenAI, 2024) for query generation, choosing a capable model to test under favorable conditions for quality-focused baselines.

Retriever Architecture. We fine-tune Contriever (Izacard et al., 2022) and RetroMAE (Xiao et al., 2022) to validate generalization across pretraining objectives. Training uses batch size 128, early stopping on dev NDCG@10.

Document: *Results-Based Accountability (RBA) is a disciplined way of thinking and taking action that communities can use to improve the lives of children, youth, families, adults and the community as a whole. RBA is also used by organizations to improve the performance of their programs...*

Supervised Query	Few-shot (InPars-GBQ, 3 queries)
what is rba	1. What is RBA and how does it help communities? 2. What is RBA and how is it used to improve community well-being? 3. What is RBA and how does it benefit communities? ↪ All ask *the same question* with minor variations

Zero-shot Diverse (Ours, 20 queries)	
1. What is Results-Based Accountability (RBA)?	11. Community impact
2. What are the main goals of RBA?	12. Conditions of well-being
3. What types of communities can benefit from RBA?	13. Performance improvement
4. How does RBA improve community well-being?	14. RBA is used to enhance community well-being.
5. How can organizations implement RBA in their programs?	15. RBA helps organizations improve program performance.
6. How do leaders measure community impact using RBA?	16. Leaders work collectively to achieve community impact.
7. Why is RBA important for improving lives?	17. Which groups can utilize RBA for improvement?
8. Why do communities need to focus on conditions of well-being?	18. Is it true that RBA focuses on children and families?
9. When should a community start using RBA?	19. How does RBA compare to traditional methods?
10. If a community adopts RBA, what changes can be expected?	20. What are the differences between community impact and program performance?

↪ Diverse formats: What/How/Why/When/If questions, keywords, statements, comparisons

Table 1: Comparison of queries generated by different methods for the first document in MS MARCO. Supervised queries are short and keyword-focused. Few-shot methods produce paraphrases of a single pattern. Our zero-shot diverse method generates semantically varied queries covering different aspects and formats.

Baselines. We compare against representative query synthesis methods. **Supervised** uses human-annotated MS MARCO queries as a reference. **InPars-GBQ** (Bonifacio et al., 2022) employs 3-shot prompting with the Guided-by-Bad-Questions variant. **DRAGON-S** (Lin et al., 2023) uses sentence-level cropping with cross-encoder reranking. **SAP** (Thakur et al., 2024) applies 5-shot summarize-then-ask prompting. **DRAMA** (Ma et al., 2025) leverages LLM-augmented data to train dense retrievers.

Evaluation Benchmarks. We evaluate on four categories using NDCG@10: (1) **TREC-DL** (Deep Learning 2019/2020 tracks) for in-domain evaluation on MS MARCO passages; (2) **BEIR** (Thakur et al., 2021) with 14 datasets for standard OOD evaluation covering factoid QA, argument retrieval, and scientific domains; (3) **BRIGHT** (Su et al., 2024) with 12 datasets for reasoning-intensive retrieval requiring complex inference; (4) **Multi-hop** including 2WikiMultihopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), HotpotQA (Yang et al., 2018), and NovelHopQA (Gupta et al., 2025), which require aggregating evidence across multiple documents.

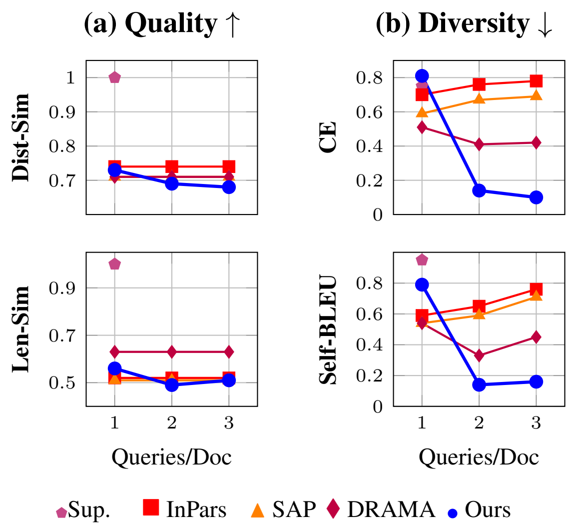

Figure 4: Quality and Diversity metrics as the number of queries per document increases. **(a) Quality** (↑): Dist-Sim and Len-Sim measure similarity to human-annotated queries, where higher values indicate more human-like quality. **(b) Diversity** (↓): CE and Self-BLEU measure query similarity, where lower values indicate higher diversity. Few-shot methods become *less diverse* (values increase), while our method becomes dramatically *more diverse* (scores drop from ∼0.8 to ∼0.1). DRAGON-S is excluded as it uses a teacher model to rerank document order, making quality and diversity metrics not directly comparable. Full data in Appendix Table 11.

5 Results

We first compare our zero-shot diverse method against few-shot baselines to establish that diversity matters, then analyze *which* tasks benefit most.

Figure 4 shows the quality-diversity trade-off. Our method produces slightly lower quality (Dist-Sim, Len-Sim) but dramatically higher diversity

Method	Q/Doc	Contriever TREC-DL (2)	Contriever BEIR (14)	Contriever BRIGHT (12)	Contriever Multi-hop (4)	RetroMAE TREC-DL (2)	RetroMAE BEIR (14)	RetroMAE BRIGHT (12)	RetroMAE Multi-hop (4)
Pretrained	-	41.68	28.73	3.84	39.70	11.52	14.27	5.20	18.45
Supervised	1	**57.52**	40.39	7.11	50.41	56.26	38.47	6.33	51.15
Non-LLM Data Augmentation									
DRAGON-S	1	4.10	7.83	1.52	2.37	3.90	6.73	1.39	2.22
DRAGON-S†	2	4.80	7.29	1.51	2.51	5.11	6.26	1.48	1.71
DRAGON-S†	3	3.88	8.31	1.64	3.27	3.20	6.09	1.60	1.34
Few-shot LLM Query Synthesis									
InPars-GBQ	1	56.33	41.31	7.84	52.72	56.07	39.02	6.98	52.39
InPars-GBQ†	2	55.31	41.26	8.44	52.58	**56.71**	39.73	7.22	53.50
InPars-GBQ†	3	54.35	41.15	8.22	53.26	56.08	39.89	7.82	53.79
SAP	1	53.53	41.37	9.49	52.58	54.53	39.69	7.43	54.85
SAP†	2	53.08	**41.50**	8.95	53.32	54.12	**40.20**	**8.63**	55.84
SAP†	3	53.42	41.41	9.00	52.78	54.30	40.04	7.94	53.96
DRAMA	1	49.61	37.36	9.43	50.72	52.12	35.38	6.42	49.42
DRAMA†	2	50.10	37.75	**9.83**	51.44	51.50	34.32	6.29	48.74
DRAMA†	3	48.14	37.47	9.50	51.41	51.71	34.69	6.06	48.97
Zero-shot LLM Query Synthesis (Ours)									
Ours	1	52.94	41.43	8.57	52.77	54.00	39.31	7.72	54.56
Ours	2	53.97	40.94	8.95	54.20	52.66	39.14	8.07	**58.50**
Ours	3	53.34	40.82	9.12	**55.22**	53.89	38.61	7.69	57.09
Ours with 8k Documents (10% of baselines)									
Ours (8k)	2	50.33	40.72	9.50	54.34	51.26	38.97	7.40	56.91
Ours (8k)	5	48.86	40.70	9.14	54.40	49.37	38.85	8.02	56.35

Table 2: Main results (NDCG@10) comparing Contriever and RetroMAE backbone retrievers. **Pretrained**: original checkpoints without MS MARCO fine-tuning, serving as initialization for all methods. **Supervised**: fine-tuned from Pretrained using human-annotated MS MARCO queries. Numbers in parentheses indicate dataset counts. All methods except "Ours (8k)" use 80k MS MARCO documents; "Ours (8k)" uses only 8k documents (10% of baselines). Both backbones show consistent patterns: our method achieves competitive or best Multi-hop performance while DRAGON-S fails catastrophically. Best in **bold**, second-best underlined. †Extended by us; original methods generate only one query per document.

Method	Q/Doc	Contriever Novel	Contriever Hotpot	Contriever MuSiQue	Contriever 2Wiki	Contriever Avg	RetroMAE Novel	RetroMAE Hotpot	RetroMAE MuSiQue	RetroMAE 2Wiki	RetroMAE Avg
Pretrained	-	37.84	41.01	32.11	47.85	39.70	21.20	20.88	12.55	19.18	18.45
Supervised	1	54.53	52.01	33.72	61.39	50.41	58.61	48.92	33.42	63.65	51.15
DRAGON-S	1	2.34	1.21	2.05	3.88	2.37	1.87	1.05	1.63	4.34	2.22
DRAGON-S†	2	1.93	1.13	1.84	5.14	2.51	1.67	0.75	1.37	3.04	1.71
DRAGON-S†	3	2.48	1.81	2.31	6.49	3.27	1.64	0.59	1.12	1.99	1.34
InPars-GBQ	1	56.49	53.85	35.62	64.94	52.73	61.38	49.91	33.95	64.31	52.39
InPars-GBQ†	2	55.77	53.82	35.89	64.83	52.58	61.63	51.95	34.25	66.19	53.50
InPars-GBQ†	3	56.99	53.78	36.27	66.00	53.26	62.48	52.94	34.25	65.50	53.79
SAP	1	57.41	55.64	35.74	61.52	52.58	66.09	53.87	35.51	63.93	54.85
SAP†	2	58.91	56.35	35.55	62.45	53.31	68.37	55.37	35.19	64.43	55.84
SAP†	3	57.22	55.93	35.70	62.27	52.78	66.60	53.15	34.71	61.38	53.96
DRAMA	1	50.12	53.61	35.86	63.30	50.72	49.34	51.36	33.42	63.55	49.42
DRAMA†	2	53.45	53.42	35.51	63.36	51.44	48.48	50.40	32.99	63.08	48.74
DRAMA†	3	53.35	53.96	35.40	62.94	51.41	49.20	50.37	33.20	63.10	48.97
Ours	1	55.08	55.58	35.39	65.02	52.77	64.20	54.38	34.64	65.03	54.56
Ours	2	60.20	55.49	36.60	64.50	54.20	**72.81**	55.42	**38.60**	67.18	**58.50**
Ours	3	**62.49**	55.62	36.54	**66.24**	**55.22**	72.25	53.29	37.38	65.42	57.09
Ours with 8k Documents (10% of baselines)											
Ours (8k)	2	58.88	55.40	**37.05**	66.03	54.34	64.52	**56.91**	38.24	**67.98**	56.91
Ours (8k)	5	62.14	**56.55**	35.72	63.21	54.41	68.99	55.96	36.50	63.93	56.35

Table 3: Detailed multi-hop retrieval results (NDCG@10) comparing Contriever and RetroMAE. 2Wiki=2WikiMultihopQA, Hotpot=HotpotQA, Novel=NovelHopQA. †Extended by us.

(CE drops from 0.50-0.81 to 0.10-0.14). Table 2 presents retrieval results:

Our method excels on reasoning tasks. Few-shot baselines produce homogeneous queries

(CE=0.50–0.78), while our method generates truly diverse queries (CE<0.15). Despite simpler prompting, we achieve the best multi-hop performance (55.22 vs. 53.26 for InPars-GBQ) while maintaining competitive BEIR (41.43). DRAGON-S's catastrophic failure (Multi-hop: 2.37) confirms that data quantity alone is insufficient without query diversity.

Multi-hop benefits most; NovelHopQA leads. Table 3 shows our method achieves the best average across all four multi-hop datasets, with largest gains on NovelHopQA (+5.5 over InPars-GBQ). Figure 5 visualizes this pattern: diverse training consistently outperforms paraphrase training on multi-hop tasks. Why does NovelHopQA benefit most? Table 4 shows it uniquely combines long contexts (∼2,336 chars) with multi-hop reasoning, requiring varied query formulations. 2WikiMultihopQA and HotpotQA can often be solved via single-hop shortcuts (Trivedi et al., 2022), leaving limited room for improvement. These observations motivate the central question of our analysis: *what property of tasks determines whether diversity helps?*

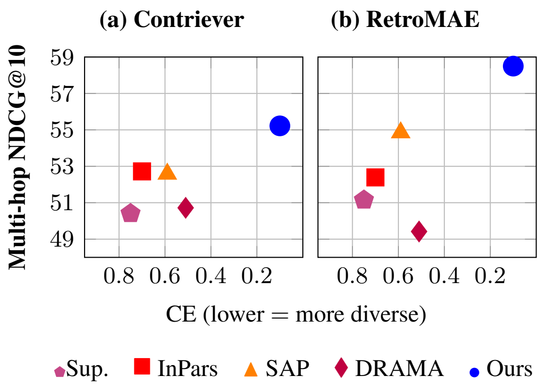

Figure 5: Query diversity correlates with multi-hop retrieval performance. X-axis: CE (cross-encoder paraphrase ratio), where lower = more diverse. Y-axis: NDCG@10 on Multi-hop benchmark. Our method with the lowest CE achieves the best performance on both retrievers.

6 Analysis

The results show diversity benefits multi-hop tasks but not uniformly. This section investigates *why*: we hypothesize that query complexity is the key factor, and validate this through controlled experiments.

Dataset	H-M Gap	Hops	Context
NovelHopQA	–	1–4	Long
HotpotQA	9.6	2	Short
MuSiQue	28.2	2–4	Short
2WikiMultihopQA	3.7	2,4	Short

Table 4: Multi-hop dataset difficulty characteristics. H-M Gap: Human-Machine Answer F1 gap from Trivedi et al. (2022) (higher = harder). While NovelHopQA lacks H-M Gap data, it is expected to be the most challenging: Gupta et al. (2025) report that even frontier models (o1, Gemini 2.5 Pro) drop below 80% accuracy on the hardest tasks (4-hop at long context), demonstrating that current LLMs struggle when hop depth and context length increase jointly.

6.1 Varying Diversity Levels

To rigorously test whether query complexity determines diversity benefit, we systematically vary diversity through two dimensions: (1) **Prompting strategy**, comparing Paraphrase (same question, varied wording) vs. Diverse (different query formats) at Q/Doc ∈ {5, 10, 20}; (2) **Query quantity**, scaling Q/Doc from 1 to 20 with diverse prompting. This yields 14 experimental conditions across 2 model architectures.

Variant	Q/Doc	Novel	Hotpot	MuSiQue	2Wiki	Avg
Paraphrase	5	52.26	52.58	35.21	**66.34**	51.60
Diverse	5	**62.14**	**56.55**	**35.72**	63.21	**54.40**
Paraphrase	10	55.97	52.74	**35.57**	**65.84**	52.53
Diverse	10	**63.29**	**55.61**	34.67	61.96	**53.88**
Paraphrase	20	53.93	51.57	**34.77**	**63.97**	51.06
Diverse	20	**64.37**	**54.93**	32.75	58.01	**52.52**

(a) Contriever

Variant	Q/Doc	Novel	Hotpot	MuSiQue	2Wiki	Avg
Paraphrase	5	59.33	50.25	33.36	**64.96**	51.97
Diverse	5	**68.99**	**55.96**	**36.50**	63.93	**56.35**
Paraphrase	10	59.64	51.51	32.89	**64.06**	52.02
Diverse	10	**71.17**	**54.50**	**35.57**	62.40	**55.91**
Paraphrase	20	60.73	**50.82**	**32.83**	**63.87**	52.06
Diverse	20	**69.91**	50.61	32.78	56.64	**52.49**

(b) RetroMAE

Table 5: NDCG@10 (%) on multi-hop datasets comparing Paraphrase (CE≈0.55) vs. Diverse (CE≈0.04) training at different Q/Doc ratios on 8k documents. Diverse consistently outperforms on Novel (+9.88 to +10.44 for Contriever; +9.17 to +11.53 for RetroMAE) across all Q/Doc settings. 2Wiki consistently favors Paraphrase, while Hotpot and MuSiQue show mixed patterns. Full results across all benchmarks are in Table 14.

Results. Table 5 shows Diverse outperforms Paraphrase on NovelHopQA (+9.88/+9.66) and HotpotQA (+3.97/+5.71), while 2WikiMultihopQA slightly favors Paraphrase. This heterogeneity suggests diversity benefit varies with query complexity.

What Drives Diversity Benefit? We hypothesize that *query complexity* explains this heterogeneity. We measure complexity using content words (CW), defined as unique non-stopwords per query:

$$\text{CW}(q) = |\{w \in \text{tokenize}(q) : w \notin \mathcal{S} \land |w| > 1\}| \quad (1)$$

where $\mathcal{S}$ is a standard stopword list. Higher CW indicates more entities and relationships that must be jointly satisfied. Such queries admit more paraphrase possibilities; training on diverse formulations enables better generalization. Table 7 shows all 14 conditions yield strong correlation between CW and diversity benefit ($r \geq 0.95$, 12/14 with $p < 0.05$). NovelHopQA (CW=11.64) benefits most; 2WikiMultihopQA (CW=6.34) shows minimal benefit.

Q/Doc	Novel	Hotpot	MuSiQue	2Wiki
1	50.67	53.33	37.01	**68.32**
2	58.88	55.40	**37.05**	66.03
5	62.14	**56.55**	35.72	63.21
10	63.29	55.61	34.67	61.96
20	**64.37**	54.93	32.75	58.01

(a) Contriever

Q/Doc	Novel	Hotpot	MuSiQue	2Wiki
1	56.11	52.10	33.12	64.90
2	64.52	**56.91**	**38.24**	**67.98**
5	68.99	55.96	36.50	63.93
10	**71.17**	54.50	35.57	62.40
20	69.91	50.61	32.78	56.64

(b) RetroMAE

Table 6: NDCG@10 (%) on four multi-hop datasets when scaling diverse queries per document (8k documents). Datasets show different optimal Q/Doc ratios: Novel benefits from more diverse queries (peaking at Q/Doc=10-20), while 2Wiki performs best with fewer queries (Q/Doc=1-2). This heterogeneity suggests that optimal diversity depends on dataset characteristics: datasets requiring diverse query formulations benefit more from additional diverse queries. Full results across all benchmarks are in Table 18.

Actionable Thresholds. Table 8 shows positive rates (proportion of conditions where diversity helps): NovelHopQA (CW=11.64) benefits in 100% of conditions, HotpotQA (CW=8.60) in 86%, MuSiQue (CW=8.64) in 43%, and 2WikiMultihopQA (CW=6.34) in only 7%. Linear regression across all 56 data points yields threshold CW=7.9 (r=0.89, p<0.0001). This suggests: **CW < 7** avoid diversity; **CW > 10** use diversity; **CW 7–10** test empirically.

Comparison	ΔNDCG@10 (%) Novel	Hotpot	MuSiQue	2Wiki	Pearson r	p
Content Words (CW)	*11.64*	*8.60*	*8.64*	*6.34*		
Diverse − Paraphrase						
Q/Doc=5	+9.9	+4.0	+0.5	−3.1	0.96	0.035*
Q/Doc=10	+7.3	+2.9	−0.9	−3.9	0.95	0.054
Q/Doc=20	+10.4	+3.4	−2.0	−6.0	0.95	0.052
Q/Doc=k − Q/Doc=1						
k=2	+8.2	+2.1	+0.0	−2.3	0.97	0.030*
k=5	+11.5	+3.2	−1.3	−5.1	0.96	0.040*
k=10	+12.6	+3.0	−2.3	−6.4	0.96	0.037*
k=20	+13.7	+1.6	−4.3	−10.3	0.97	0.032*

(a) Contriever

Comparison	ΔNDCG@10 (%) Novel	Hotpot	MuSiQue	2Wiki	Pearson r	p
Content Words (CW)	*11.64*	*8.60*	*8.64*	*6.34*		
Diverse − Paraphrase						
Q/Doc=5	+9.7	+5.7	+3.1	−1.0	0.96	0.036*
Q/Doc=10	+11.5	+3.0	+2.7	−1.7	0.99	0.008*
Q/Doc=20	+9.2	−0.2	−0.0	−7.2	1.00	<0.001*
Q/Doc=k − Q/Doc=1						
k=2	+8.4	+4.8	+5.1	+3.1	0.99	0.007*
k=5	+12.9	+3.9	+3.4	−1.0	0.99	0.010*
k=10	+15.1	+2.4	+2.5	−2.5	0.98	0.021*
k=20	+13.8	−1.5	−0.3	−8.3	0.99	0.010*

(b) RetroMAE

Table 7: Content words (CW) and diversity benefit (ΔNDCG@10) on four multi-hop datasets. For each experimental condition, we compute the Pearson correlation between the four CW values and the corresponding four ΔNDCG@10 values (n=4). Two types of ΔNDCG@10 comparisons: (1) Diverse − Paraphrase training at fixed Q/Doc, and (2) Q/Doc=k − Q/Doc=1 with diverse training. All 14 conditions show strong positive correlation ($r \geq 0.95$), with 12/14 significant (*$p < 0.05$). Data from 8k MS MARCO documents.

CW Range	Positive Rate	Diversity
CW < 7	7% (1/14)	Avoid
CW 7–10	43–86%	Test
CW > 10	100% (14/14)	Recommend

Table 8: CW-based diversity recommendations for multi-hop QA tasks. Positive rate indicates the proportion of 14 experimental conditions (7 Contriever + 7 RetroMAE) where diversity training improves performance (ΔNDCG@10 > 0). Linear regression across 56 data points (4 datasets × 14 conditions) yields threshold CW=7.9 (where Δ=0). These recommendations apply only to reasoning-intensive tasks where CW-diversity correlation holds (r=0.89, p<0.0001).

The Complexity-Diversity Principle (CDP). We formalize our findings as the **Complexity-Diversity Principle (CDP)**: *query complexity determines whether diversity helps or hurts*. Complex queries can be expressed in many semantically distinct ways, and training on diverse formulations

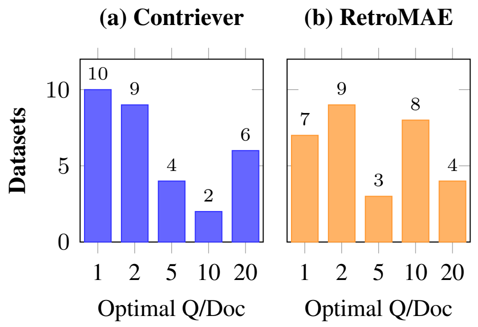

Figure 6: Distribution of optimal Q/Doc ratios across 31 datasets. The optimal ratio varies from 1 to 20, with no universally optimal value.

helps learn this rich space. Simple queries have limited paraphrase possibilities; excessive diversity introduces noise. Standard retrieval tasks have uniform query patterns where similar training suffices, but reasoning-intensive tasks require handling diverse formulations.

Diversity as Regularization. Error analysis (Appendix F) reveals 87.2% of M=3's improvements over M=1 on NovelHopQA are "same-book-wrong-passage" corrections; M=1 overfits to superficial features (e.g., character names), while M=3 learns fine-grained semantic distinctions. This explains why Multi-hop improves from M=1 (52.77) to M=3 (55.22), while BEIR slightly decreases (41.43→40.82). Figure 6 confirms: 68% of datasets benefit from higher diversity (Q/Doc>1), even higher (77%) for RetroMAE.

6.2 Document Quantity vs. Query Diversity

The CDP explains *when* diversity helps; practitioners also need to know: can diversity substitute for document quantity? Fewer documents with more queries per document reduces LLM cost while maintaining training pairs.

Trading Documents for Diversity. Figure 7 shows the trade-off across CW levels: (1) **High-CW (CW > 10)**: NovelHopQA improves from 64.2 to 71.2 as documents decrease from 80k to 8k, indicating that diversity effectively substitutes for quantity. (2) **Mid-CW (CW 7–10)**: HotpotQA and MuSiQue show mixed patterns. (3) **Low-CW (CW < 7)**: 2WikiMQA consistently degrades with fewer documents. Statistical analysis (n=8) confirms CW predicts this trade-off (r=−0.88, p<0.01).

Practical Implications. For high-CW tasks, practitioners can achieve state-of-the-art with only

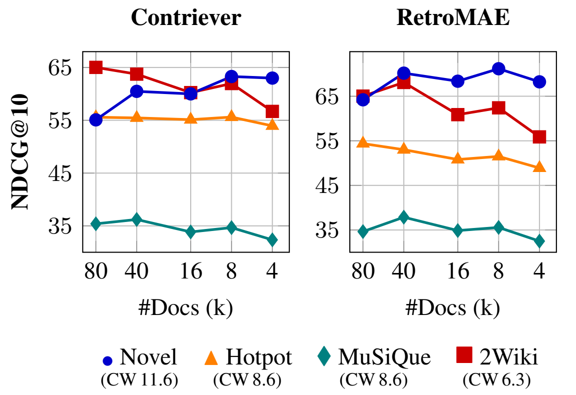

Figure 7: Document quantity vs. query diversity trade-off on multi-hop QA tasks. Legend shows dataset name and CW value; colors progress from red (low CW) to blue (high CW). With fixed 80k training pairs, we vary configurations from 80k×1 to 4k×20. High-CW datasets (NovelHopQA) show upward trends as documents decrease, while low-CW datasets (2WikiMQA) show downward trends, confirming that high-CW tasks benefit more from query diversity than from document coverage.

10% of documents. Configurations like 40k×2 or 16k×5 achieve 50–80% cost reduction with minimal degradation, reinforcing that high-complexity tasks benefit from diversity more than document coverage.

7 Conclusion

We address conflicting evidence on query diversity in three stages. First, we design Q-D metrics to quantify diversity's impact on retriever training, making the problem measurable. Second, through experiments on 4 benchmark types (31 datasets), we find query diversity especially benefits multi-hop retrieval. Third, deep analysis on multi-hop data reveals the **Complexity-Diversity Principle (CDP)**: *query complexity determines optimal diversity* (r≥0.95). Guided by CDP, we propose zero-shot multi-query synthesis, achieving state-of-the-art on multi-hop tasks.

Practical Guidelines. CW > 10: use diversity (100% positive rate). CW < 7: avoid diversity (7% positive rate). CW 7–10: test empirically.

Broader Impact. Rather than reporting "diversity sometimes helps," we identify the underlying factor and provide quantitative predictions. Future work includes automatic complexity estimation and extension to multilingual retrieval.

Limitations

We validate CDP on two dense retriever architectures; extending to sparse retrievers (SPLADE) and late-interaction models (ColBERT) would strengthen generalizability. Our CW metric captures lexical complexity; syntactic or semantic measures (dependency depth, entity count) may provide additional insights. We control diversity via prompt instructions; alternative mechanisms warrant exploration. Finally, we focus on English; CDP may vary across languages with different morphological characteristics.

Ethics Statement

Our work uses publicly available datasets (MS MARCO, BEIR, BRIGHT) and commercial LLM APIs (GPT-4o-mini). We do not collect or process personal data. The synthetic queries are generated for research purposes and may contain biases inherent in the LLM. We encourage responsible use of synthetic data generation techniques.

References

Luiz Henrique 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*, pages 2622–2631.

Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B Hall, and Ming-Wei Chang. 2023. Promptagator: Few-shot dense retrieval from 8 examples. In *International Conference on Learning Representations*.

Abhay Gupta, Michael Lu, Kevin Zhu, Sean O'Brien, and Vasu Sharma. 2025. NovelHopQA: Diagnosing multi-hop reasoning failures in long narrative contexts. *arXiv preprint arXiv:2506.02000*.

Daniel Gwon, Nour Jedidi, and Jimmy Lin. 2025. Study on LLMs for promptagator-style dense retriever training. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. In *Transactions on Machine Learning Research*.

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.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. 2023. How to train your DRAGON: Diverse augmentation towards generalizable dense retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6385–6400.

Xueguang Ma, Luyu Gao, Jakub Thorne, Wenhu Chen, Jimmy Lin, and Michiel de Jong. 2025. DRAMA: Diverse augmentation from large language models to smaller dense retrievers. *arXiv preprint arXiv:2502.18460*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan Ö. Arik, Danqi Chen, and Tao Yu. 2024. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*.

Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, and Daniel Cer. 2024. Leveraging LLMs for synthesizing training data across many languages in multilingual dense retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 7749–7763.

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*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. In *arXiv preprint arXiv:1807.03748*.

Shitao Xiao, Zheng Liu, Yingxia Shao, and Zhao Cao. 2022. RetroMAE: Pre-training retrieval-oriented language models via masked auto-encoder. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 538–548.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

A Algorithm

Algorithm 1 provides the complete pseudocode for our zero-shot multi-query generation approach, including the prompt tuning phase where Q-D metrics guide prompt selection.

Algorithm 1 Zero-shot Multi-Query Generation with Q-D Guided Prompt Tuning

Require: Corpus $\mathcal{D}$, queries per doc M, LLM $\mathcal{L}$, target task (in-domain/OOD)
Ensure: Training set $\mathcal{T}$

// Phase 1: Prompt Tuning with Q-D Metrics
$\mathcal{D}_{\text{sample}} \leftarrow$ sample n documents from $\mathcal{D}$ {e.g., n=100}
$\mathcal{P} \leftarrow \{p_{\text{paraphrase}}, p_{\text{diverse}}, ...\}$ {Candidate prompts}
for each prompt $p \in \mathcal{P}$ **do**
  Generate queries $\mathcal{Q}_p$ using p on $\mathcal{D}_{\text{sample}}$
  Compute Q-D metrics: CE$_p$, Self-BLEU$_p$
end for
Select $p^* \leftarrow$ prompt matching target diversity {See Fig. 2}
  In-domain: CE > 0.5, Self-BLEU > 0.5
  OOD: CE < 0.5, Self-BLEU < 0.5
// Phase 2: Full-scale Query Generation
$\mathcal{T} \leftarrow \emptyset$
for each document $d \in \mathcal{D}$ **do**
  $\{q_1, ..., q_M\} \leftarrow \mathcal{L}(p^*(d, M), \text{temp} = 0)$ {Single call}
  for $i = 1$ to M **do**
    $\mathcal{T} \leftarrow \mathcal{T} \cup \{(q_i, d)\}$
  end for
end for
Train retriever f_θ on $\mathcal{T}$ with InfoNCE loss

B Query Generation Examples

Table 9 shows the 20 queries generated by the zero-shot paraphrase method for the same document used in Table 1. Unlike our diverse method, paraphrase queries all follow the same question pattern with only surface-level variations.

C Implementation Details

Training Configuration We train all retriever models using the sentence-transformers library. We use AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-8}$, and weight decay of 0.01. The learning rate is set to 2×10^{-6} for Contriever and 5×10^{-6} for RetroMAE, selected via grid search using supervised human-annotated data on the MS MARCO development set, then applied uniformly to all methods. We use cosine learning rate decay without warmup. The batch size is 128, and models are trained for up to 100 epochs with gradient clipping (max norm 1.0). We use InfoNCE loss (van den Oord et al., 2018) with scale factor 20.0 as the loss function, and cosine similarity for scoring. Training employs FP16 mixed precision with gradient checkpointing enabled. Checkpoints are saved and evaluated every 500 steps, with model selection based on NDCG@10 on the MS MARCO development set. We use the official Contriever (`facebook/contriever`) and RetroMAE (`Shitao/RetroMAE`) checkpoints as initialization.

Temperature Setting for LLM-based Query Generation For our zero-shot method, we use temperature=0 to ensure reproducibility, as our prompt explicitly instructs the LLM to generate multiple diverse queries in a single API call. For few-shot baselines (InPars-GBQ, SAP, DRAMA), we use temperature=0.7 when generating multiple queries per document, since few-shot prompts typically produce a single query per call and require multiple sampling to obtain M queries.

Learning Rate Selection We conduct a learning rate sweep using supervised human-annotated MS MARCO data, selecting from $\{2 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, 2 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}\}$. Table 10 shows the best learning rate at different training checkpoints. The selected learning rate is then applied uniformly to all data synthesis methods.

Document: *Results-Based Accountability (RBA) is a disciplined way of thinking and taking action that communities can use to improve the lives of children, youth, families, adults and the community as a whole. RBA is also used by organizations to improve the performance of their programs...*

Zero-shot Paraphrase (20 queries)

1. What is RBA and how does it benefit communities?
2. How can communities utilize RBA to enhance lives?
3. In what ways does RBA contribute to community improvement?
4. What are the main advantages of implementing RBA?
5. How does RBA help organizations improve effectiveness?
6. What role does community impact play in RBA?
7. How does RBA address the needs of children and families?
8. What is the purpose of using RBA in community initiatives?
9. How can leaders in a community apply RBA?
10. What outcomes can be expected from implementing RBA?
11. How does RBA facilitate collective action among leaders?
12. What is the significance of community well-being in RBA?
13. How does RBA measure community success?
14. What strategies can communities adopt using RBA?
15. How does RBA support safe and clean neighborhoods?
16. What is the impact of RBA on program performance?
17. How can RBA create better job opportunities?
18. What is the relationship between RBA and leadership?
19. How does RBA influence readiness of children for school?
20. What are the key components of community impact in RBA?

↪ All queries follow the same "What/How does RBA..." pattern with surface-level variations only

Table 9: Queries generated by the zero-shot paraphrase method for the first document in MS MARCO. Unlike our diverse method (Table 1), all queries follow similar question patterns, lacking format diversity (no keywords, statements, or comparisons).

Retriever	Checkpoint	Best LR	NDCG@10
Contriever	10k steps	1×10^{-5}	92.86
	20k steps	2×10^{-6}	92.83
	30k steps	2×10^{-6}	92.87
	Overall best	2×10^{-6}	**92.87**
RetroMAE	10k steps	1×10^{-5}	91.52
	20k steps	5×10^{-6}	91.87
	30k steps	5×10^{-6}	91.73
	Overall best	5×10^{-6}	**91.87**

Table 10: Learning rate sensitivity analysis on the MS MARCO development set. We report the best learning rate at different training checkpoints (10k, 20k, 30k steps) and the overall best throughout 100-epoch training. The overall best learning rates (2×10^{-6} for Contriever, 5×10^{-6} for RetroMAE) are used for all experiments.

Method	Q/Doc	Quality		Diversity	
		Dist-Sim↑	Len-Sim↑	CE↓	Self-BLEU↓
Supervised	1	1.00	1.00	0.75	0.95
InPars-GBQ	1	0.74	0.52	0.70	0.59
	2†	0.74	0.52	0.76	0.65
	3†	0.74	0.52	0.78	0.76
SAP	1	0.71	0.51	0.59	0.54
	2†	0.71	0.51	0.67	0.59
	3†	0.71	0.51	0.69	0.71
DRAMA	1	0.71	0.63	0.51	0.54
	2†	0.71	0.63	0.41	0.33
	3†	0.71	0.63	0.42	0.45
Ours	1	0.73	0.56	0.81	0.79
Ours	2	0.69	0.49	0.14	0.14
Ours	3	0.68	0.51	0.10	0.16

Table 11: Full quality and diversity metrics. Quality metrics (Dist-Sim, Len-Sim) measure similarity to human-written queries. Diversity metrics (CE, Self-BLEU) measure variation among generated queries for the same document. Our method achieves dramatically lower CE and Self-BLEU at Q/Doc>1, indicating truly diverse query generation. †Extended by us for fair comparison; original methods generate only one query per document.

D Full Quality-Diversity Metrics

Table 11 provides the complete quality and diversity metrics for all methods and configurations.

E Dataset Statistics

Table 12 provides detailed statistics for all evaluation datasets described in Section 4.

F Error Analysis on NovelHopQA

Table 13 provides detailed error analysis comparing **Q/Doc=1** and **Q/Doc=3** predictions on NovelHopQA. Figure 8 shows additional case studies demonstrating the "same-book-wrong-passage" error pattern.

G Full Diversity Ablation Results

Table 14 provides the complete results comparing Paraphrase and Diverse training across all benchmark types, complementing Table 5 in the main text. Figure 9 visualizes the Q-D metrics for these configurations.

H Full Query Scaling Results

Tables 15, 16, and 17 provide detailed query scaling results for each benchmark type, complementing Table 6 in the main text. Table 18 provides a summary across all benchmark types.

I Cost Efficiency Analysis

Table 19 provides full results for the cost efficiency experiment shown in Figure 7. All configurations use 80k total training pairs but vary the document-query ratio.

Type	Dataset	#Queries	#Docs	Query Len	Doc Len	Docs/Query	Queries/Doc
TREC-DL	trec-dl-2019	43	8,841,823	33	335	215	1
	trec-dl-2020	54	8,841,823	34	335	211	1
BEIR	arguana	1,406	8,674	1,193	1,030	1	1
	climate-fever	1,535	5,416,593	123	539	3	3
	cqadupstack	13,145	457,199	50	932	2	1
	dbpedia-entity	400	4,635,922	34	310	109	1
	fever	6,666	5,416,568	50	539	1	5
	fiqa	648	57,638	63	767	3	1
	hotpotqa	7,405	5,233,329	92	289	2	1
	nfcorpus	323	3,633	22	1,591	38	4
	nq	3,452	2,681,468	48	493	1	1
	quora	10,000	522,931	52	62	2	1
	scidocs	1,000	25,657	72	1,204	30	1
	scifact	300	5,183	90	1,499	1	1
	trec-covid	50	171,332	69	1,118	1,327	2
	webis-touche2020	49	382,545	43	1,720	45	1
BRIGHT	aops	111	188,002	320	754	5	5
	biology	103	57,359	523	330	4	1
	earth_science	116	121,249	477	338	5	1
	economics	103	50,220	740	395	8	1
	leetcode	142	413,932	1,459	1,059	2	1
	pony	112	7,894	389	260	20	52
	psychology	101	52,835	693	384	7	1
	robotics	101	61,961	2,180	291	5	1
	stackoverflow	117	107,081	1,293	1,715	4	1
	sustainable_living	108	60,792	683	344	5	1
	theoremqa_questions	194	188,002	426	754	3	2
	theoremqa_theorems	76	23,839	416	874	2	2
Multi-hop	2wikimultihopqa	12,576	125,237	68	377	2	1
	musique	2,417	48,315	102	524	3	1
	novelhopqa	4,345	4,345	138	2,336	1	1
	hotpotqa	7,405	5,233,329	92	289	2	1

Table 12: Statistics of evaluation benchmarks. Query Len and Doc Len are average character lengths. HotpotQA appears in both BEIR and Multi-hop categories as they refer to the same dataset.

Query: What is Leopold Bloom accused of being?
Book: *Ulysses*

Q/Doc=1 ✗ *"BLOOM: I have forgotten for the moment. Ah, yes! (He takes off his high grad..."*
↪ A dialogue mentioning Bloom

Q/Doc=3 ✓ *"THE CRIER: (Loudly.) Whereas Leopold Bloom of no fixed abode is a wellknown dynamitard..."*
↪ The actual accusation against Bloom

Query: Why does Elinor suggest deferring the letter?
Book: *Sense and Sensibility*

Q/Doc=1 ✗ *"Elinor said no more. She was debating within herself on the eligibility of b..."*
↪ Elinor in a different context

Q/Doc=3 ✓ *"As dinner was not to be ready in less than two hours from their arrival, Eli..."*
↪ Elinor's suggestion about the letter

Query: Who prevented Sancho from being robbed?
Book: *Don Quixote*

Q/Doc=1 ✗ *"Sancho came home in such glee and spirits that his wife noticed his happiness..."*
↪ Sancho in a different scene

Q/Doc=3 ✓ *"And now day dawned; and if the dead free-booters had scared them, their hearts were no..."*
↪ The robbery prevention scene

Query: What is Emma questioning about Mr. Knightley?
Book: *Emma*

Q/Doc=1 ✗ *"Well, I believe, if you will excuse me, Mr. Knightley, if you will not consider..."*
↪ A dialogue with Mr. Knightley

Q/Doc=3 ✓ *"Emma could not help laughing as she an-swered, 'Upon my word, I believe you know her..."*
↪ Emma questioning Knightley's understanding

Figure 8: Additional Contriever case studies from NovelHopQA. In all cases, Q/Doc=1 retrieves passages that merely mention the character name, while Q/Doc=3 correctly identifies passages containing the specific semantic content requested by the query.

Metric	Value
Total queries	4,345
Q/Doc=3 correct, **Q/Doc=1** wrong	522
Q/Doc=1 correct, **Q/Doc=3** wrong	187
Net improvement (**Q/Doc=3** − **Q/Doc=1**)	+335
*Among **Q/Doc=1** errors where **Q/Doc=3** succeeds:*	
Same book, wrong passage	455 (87.2%)
Different book	67 (12.8%)

Table 13: Error analysis comparing **Q/Doc=1** and **Q/Doc=3** on NovelHopQA (P@1) using Contriever. Statistics are computed on the **full test set** (4,345 queries across all 4 hop levels). The majority of **Q/Doc=1** errors retrieve the wrong passage from the correct book, indicating overfitting to surface features.

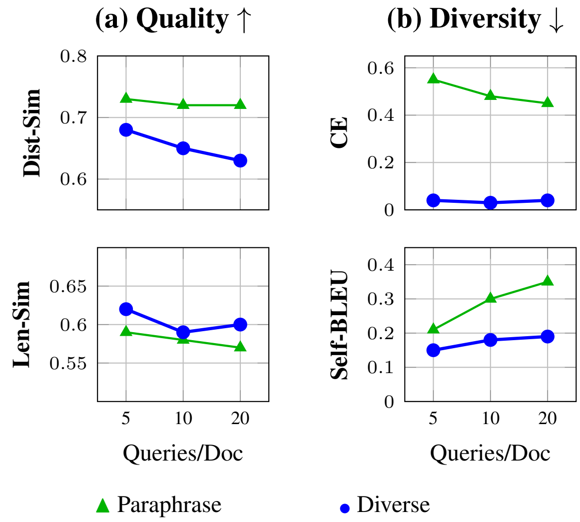

Figure 9: Quality (↑) and Diversity (↓) metrics for Paraphrase vs Diverse variants (trained on 8k documents). Higher quality values indicate more similarity to human-annotated queries; lower diversity values indicate higher diversity. Both maintain similar quality, but Diverse achieves dramatically lower CE and Self-BLEU.

Figures 10 and 11 show results across all four benchmark types (TREC-DL, BEIR, BRIGHT, Multi-hop), comparing our method with baselines that use 80k documents. The pattern differs by task type: TREC-DL and BEIR prefer document coverage, while BRIGHT and Multi-hop tolerate or benefit from reduced documents with increased query diversity.

J Cross-Condition Consistency for Threshold Estimation

Our CW threshold recommendations (Table 8) are derived from only 4 multi-hop datasets. Here we explain why the consistency across 14 experimental conditions provides reliability for these estimates.

Variant	Q/Doc	TREC-DL	BEIR	BRIGHT	Multi-hop
Paraphrase	5	**54.89**	**41.15**	7.89	51.60
Diverse	5	48.86	40.70	**9.14**	**54.40**
Paraphrase	10	**52.99**	**40.36**	8.22	52.53
Diverse	10	47.22	39.56	**9.26**	**53.88**
Paraphrase	20	**53.72**	**40.08**	8.23	51.06
Diverse	20	44.37	38.35	**9.30**	**52.52**

(a) Contriever

Variant	Q/Doc	TREC-DL	BEIR	BRIGHT	Multi-hop
Paraphrase	5	**54.34**	**39.67**	7.15	51.97
Diverse	5	49.37	38.85	**8.02**	**56.35**
Paraphrase	10	**51.98**	**38.81**	8.01	52.02
Diverse	10	46.61	37.91	**8.25**	**55.91**
Paraphrase	20	**52.30**	**38.59**	**8.00**	52.06
Diverse	20	43.27	36.46	7.86	**52.49**

(b) RetroMAE

Table 14: Effect of diversity level at different query counts, trained on 8k documents. Paraphrase generates semantically similar queries (CE≈0.50), while Diverse generates varied queries (CE≈0.04). Both backbones show consistent patterns: Paraphrase performs better on in-domain (TREC-DL) and standard OOD (BEIR), while Diverse excels on reasoning-intensive (BRIGHT) and multi-hop tasks.

Q/Doc	DL-19	DL-20
1	**52.46**	**50.55**
2	50.66	50.01
5	48.57	49.16
10	46.80	47.64
20	44.93	43.81

(a) Contriever

Q/Doc	DL-19	DL-20
1	54.04	46.99
2	**55.93**	**46.60**
5	52.76	45.98
10	49.50	43.73
20	45.86	40.68

(b) RetroMAE

Table 15: NDCG@10 (%) on TREC-DL datasets when scaling diverse queries per document (8k documents). DL-19=TREC-DL-2019, DL-20=TREC-DL-2020. In-domain performance degrades monotonically with more queries for Contriever, while RetroMAE peaks at Q/Doc=2.

The Challenge of Small Sample Size. With only 4 datasets, a single correlation coefficient could be spurious. A significant result in one experimental condition might be due to chance or specific methodological choices.

Cross-Condition Consistency as Replication. Instead of relying on a single experiment, we compute the same analysis across 14 different experimental conditions:

Q/Doc	ArguAna	C-FEVER	CQADup	DBPedia	FEVER	FiQA	NFCorpus	NQ	Quora	SCIDOCS	SciFact	T-COVID	Touche
1	44.04	21.25	**29.75**	33.39	**67.17**	24.59	31.59	29.24	**83.68**	17.29	59.48	**38.01**	**18.91**
2	50.46	26.39	29.92	33.45	66.71	**26.03**	**32.86**	**30.30**	83.19	**17.61**	64.39	36.14	17.25
5	55.18	**27.21**	29.16	**33.91**	65.39	25.92	32.55	29.28	81.95	17.23	66.02	35.12	14.35
10	**56.46**	25.31	28.19	33.61	60.93	25.34	32.10	28.22	80.90	16.88	66.06	31.66	12.57
20	56.12	23.97	27.77	33.29	55.03	24.37	32.04	24.25	80.39	16.27	**67.56**	30.88	10.09

(a) Contriever

Q/Doc	ArguAna	C-FEVER	CQADup	DBPedia	FEVER	FiQA	NFCorpus	NQ	Quora	SCIDOCS	SciFact	T-COVID	Touche
1	40.59	23.28	25.37	31.89	**71.28**	19.48	27.25	28.32	**81.82**	14.22	50.34	**44.40**	**19.94**
2	46.85	**25.15**	**26.69**	**32.18**	68.85	21.44	28.74	**29.57**	81.40	14.64	55.66	41.61	18.90
5	52.08	25.14	26.42	31.59	64.84	22.03	**28.97**	28.96	80.16	**14.76**	58.35	40.20	17.47
10	53.01	23.86	26.23	30.46	58.06	**22.77**	28.93	27.88	79.40	14.16	58.69	39.56	16.29
20	**53.22**	22.79	25.42	29.00	52.38	21.54	28.78	24.02	77.95	13.98	**60.32**	39.06	13.39

(b) RetroMAE

Table 16: NDCG@10 (%) on BEIR datasets when scaling diverse queries per document (8k documents). C-FEVER=Climate-FEVER, CQADup=CQADupStack, T-COVID=TREC-COVID, Touche=Touche-2020. Most datasets peak at Q/Doc=1-2, with SciFact and ArguAna benefiting from more queries.

Q/Doc	AOPS	Biology	Earth Sci.	Econ.	LeetCode	Pony	Psych.	Robotics	SO	Sustain.	TQA-Q	TQA-T
1	2.86	6.45	11.90	9.12	12.19	6.78	**13.43**	5.34	7.11	**7.89**	5.91	1.65
2	3.29	**9.38**	17.13	**10.70**	13.33	**10.28**	12.99	7.58	**9.63**	8.00	7.64	4.10
5	3.80	8.32	**19.13**	9.47	14.02	6.48	11.28	7.68	8.81	6.80	8.70	5.13
10	**4.68**	8.73	18.60	9.67	14.64	6.76	11.10	7.92	8.23	6.31	9.14	5.39
20	4.51	7.93	18.76	8.33	**14.49**	7.78	11.51	**8.97**	8.54	4.72	**9.67**	**6.42**

(a) Contriever

Q/Doc	AOPS	Biology	Earth Sci.	Econ.	LeetCode	Pony	Psych.	Robotics	SO	Sustain.	TQA-Q	TQA-T
1	0.93	4.81	13.52	**10.61**	12.39	**8.66**	10.75	7.11	5.81	**8.95**	2.95	1.45
2	2.78	6.52	13.92	8.00	13.65	7.12	11.03	8.06	5.65	6.70	4.05	1.35
5	3.43	7.89	**17.14**	8.86	14.80	4.86	10.61	9.29	5.04	6.51	5.34	2.43
10	**3.73**	**8.11**	16.71	8.76	**14.79**	5.05	**11.47**	**9.84**	5.00	6.18	5.34	**4.02**
20	3.26	6.67	16.32	7.54	14.64	5.00	10.93	8.48	**6.80**	5.39	**5.72**	3.54

(b) RetroMAE

Table 17: NDCG@10 (%) on BRIGHT datasets when scaling diverse queries per document (8k documents). Earth Sci.=Earth Science, Econ.=Economics, Psych.=Psychology, SO=StackOverflow, Sustain.=Sustainable Living, TQA-Q=TheoremQA Questions, TQA-T=TheoremQA Theorems. Reasoning-intensive datasets show varied optimal Q/Doc ratios, with math-related tasks (AOPS, TheoremQA) generally benefiting from more diverse queries.

Q/Doc	TREC-DL	BEIR	BRIGHT	Multi-hop
1	**51.50**	39.41	7.55	52.33
2	50.33	**40.72**	**9.50**	54.34
5	48.86	40.70	9.14	**54.40**
10	47.22	39.56	9.26	53.88
20	44.37	38.35	9.30	52.52

(a) Contriever

Q/Doc	TREC-DL	BEIR	BRIGHT	Multi-hop
1	50.52	37.88	7.33	51.56
2	**51.26**	**38.97**	7.40	**56.91**
5	49.37	38.85	8.02	56.35
10	46.61	37.91	**8.25**	55.91
20	43.27	36.46	7.86	52.49

(b) RetroMAE

Table 18: Effect of scaling diverse queries per document (8k documents). Performance peaks around Q/Doc=2-5 for most OOD benchmarks, then declines with more queries. Note that this optimal range is relative to the 8k document subset; larger corpora may require more queries per document to achieve optimal coverage. Best results per column are in **bold**.

Docs	Q/Doc	LLM Cost	TREC-DL	BEIR	BRIGHT	Multi-hop
80k	1	100%	**52.94**	**41.43**	8.57	52.77
40k	2	50%	52.04	40.89	8.89	**53.97**
16k	5	20%	48.67	41.14	**9.32**	52.30
8k	10	10%	47.22	39.56	9.26	53.88
4k	20	5%	45.66	38.18	9.14	51.49

(a) Contriever

Docs	Q/Doc	LLM Cost	TREC-DL	BEIR	BRIGHT	Multi-hop
80k	1	100%	**54.00**	39.31	7.72	54.56
40k	2	50%	53.88	**39.55**	7.73	**55.79**
16k	5	20%	50.03	39.45	7.89	54.23
8k	10	10%	46.61	37.91	**8.25**	55.91
4k	20	5%	44.44	36.86	8.05	51.88

(b) RetroMAE

Table 19: Efficiency analysis with fixed 80k training pairs. LLM cost is proportional to the number of documents (not queries per document), since multiple queries can be generated in a single API call. Reducing documents from 80k to 4k (5% LLM cost) maintains competitive OOD performance, especially on BRIGHT.

- 3 Diverse vs. Paraphrase comparisons (Q/Doc $\in \{5, 10, 20\}$)
- 4 Q/Doc scaling comparisons ($k \in \{2, 5, 10, 20\}$ vs. k=1)
- 2 model architectures (Contriever, RetroMAE)

Each condition represents a different operationalization of "high diversity vs. low diversity." If the CW-diversity relationship holds across all these variations, it is unlikely to be an artifact of any specific experimental choice.

Positive Rate as a Robust Metric. For each dataset, we compute the *positive rate*, defined as the proportion of conditions where diversity improves performance (ΔNDCG@10 > 0). This metric is robust because:

- It aggregates evidence across multiple conditions
- It is based on the sign of the effect (positive/negative), not magnitude
- It provides a probabilistic interpretation: "In X% of experimental setups, diversity helps"

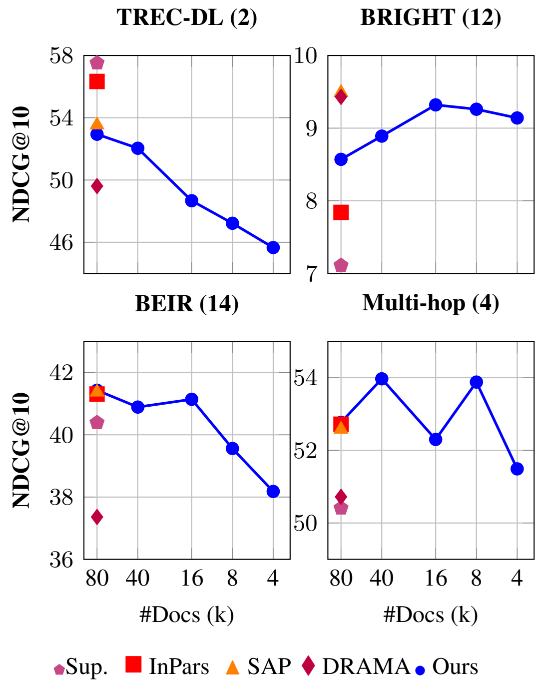

Figure 10: Cost efficiency across all benchmark types with Contriever. We vary document-query configurations from 80k×1 to 4k×20 (x-axis shows document count). Baseline methods use 80k documents. TREC-DL and BEIR prefer higher document coverage, while BRIGHT and Multi-hop tolerate or benefit from reduced documents with increased query diversity.

Results. The positive rates show a clear monotonic relationship with CW: NovelHopQA (CW=11.64) benefits in 100% of conditions, HotpotQA (CW=8.60) in 86%, MuSiQue (CW=8.64) in 43%, and 2WikiMultihopQA (CW=6.34) in only 7%. Linear regression across all 56 data points (4 datasets × 14 conditions) yields r=0.89 (p<0.0001), confirming the strong relationship.

Limitations. This approach increases *internal validity* (confidence that the relationship is real) but does not address *external validity* (generalization to other datasets). The thresholds are validated only for reasoning-intensive multi-hop QA tasks where CW-diversity correlation holds. For other

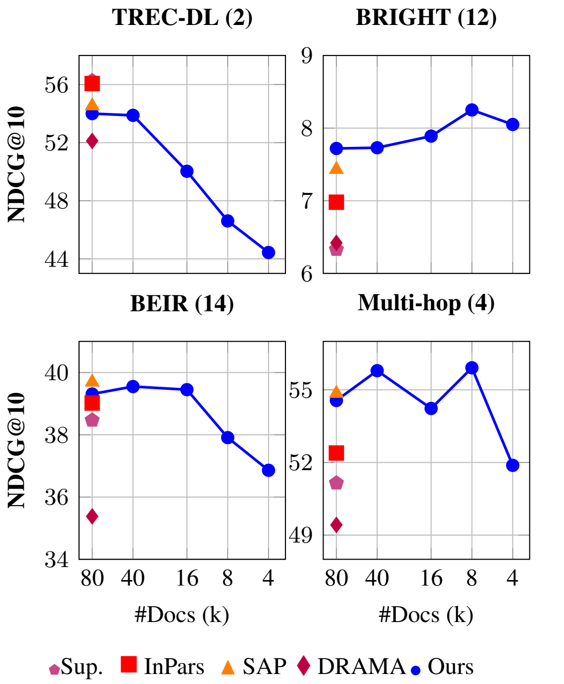

Figure 11: Cost efficiency with fixed 80k training pairs (RetroMAE backbone). Configuration and legend same as Figure 10. RetroMAE shows similar patterns: our method maintains competitive performance with fewer documents, particularly excelling on Multi-hop where 8k documents (10% cost) achieves 55.91 NDCG@10, outperforming all 80k-document baselines.

task types (e.g., BEIR, BRIGHT), the relationship may not apply.

K Limitations of Few-shot Methods

Existing methods like InPars (Bonifacio et al., 2022) and SAP (Thakur et al., 2024) employ few-shot prompting with carefully selected examples. While few-shot prompting can ensure high-quality queries that closely match target distributions, it has two critical limitations:

High Cost. Few-shot methods require M separate LLM calls to generate M queries per document, and depend on powerful LLMs to follow complex few-shot patterns.

Limited Diversity. LLMs tend to mimic the patterns of few-shot examples (syntax, length, question type), causing the output distribution to be anchored around these examples. Even with high temperature sampling (>0.7), the generated queries remain clustered around the exemplar patterns, producing only surface-level lexical variations rather than true semantic diversity. This results in high paraphrase ratios (CE>0.5), as shown in Table 1: few-shot queries for the same document are largely paraphrases of a single underlying pattern.

L Full Prompt Templates

The core prompt structure is shown in Figure 3 (main text). Here we provide the complete templates with all formatting details.

Paraphrase Prompt

Your task is to generate {M} paraphrase queries based on the document(s).

– Identify **ONE main question** the document(s) answer
– Then rephrase it {M} different ways

All queries must ask the SAME question with DIFFERENT wording.

Document(s): {document}

Generate {M} queries: 1.

Diverse Prompt for Multi-hop Tasks

Your task is to generate {M} independent queries based on the document(s).

You MUST generate queries in these specific formats:
– **What**... questions (factual)
– **How**... questions (procedural)
– **Why**... questions (causal)
– **When/If**... questions (conditional)
– **Keyword** queries (2-5 words, no question mark)
– **Statement/claim** format (e.g., "X is used for Y")
– **Which/Is it true**... questions
– **Comparison** or contrast questions

Each query must target different information from the document.

Document(s): {document}

Generate {M} queries: 1.

Figure 12: Full prompt templates. M denotes Q/Doc (the number of queries per document). To control experimental variables, when $M < 20$, we take the first M generated queries in order.