BGCT: BM25-Guided Contrastive Training for Generative Retrieval

Ping-Hsun Wu
Department of Data Science
Soochow University

Jheng-Long Wu
Department of Data Science
Soochow University

Abstract

Generative Retrieval (GR) encodes a corpus into model parameters and retrieves documents by generating their identifiers, but underperforms on lexical queries where BM25 excels. We propose BGCT (**B**M25-**G**uided **C**ontrastive **T**raining), combining BM25 Hard Negative Mining with BM25 Knowledge Distillation, a KL-divergence loss on lexically-overlapping queries. On NQ320K, GR and BM25 candidate sets are highly complementary (mean Jaccard $= 0.0835$; theoretical union R@1 ceiling 78.16%). Contrary to this promise, BGCT does not surpass our locally reproduced TSGEN baseline (67.57% R@1; the original paper reports 70.82%): BGCT-HN reaches 63.60% and BGCT-Full 55.64%. An isolated KD-only ablation (59.28%) shows KD is independently harmful: its uniform-smoothing target converges, but toward uniformity rather than confident top-1 generation. A sweep over six KD weights shows R@1 decreases monotonically with the weight, so the flaw is directional rather than a tuning issue, and paired McNemar tests show the KD and HN regressions are significant ($p < 10^{-30}$). We report these findings as a cautionary result: naively porting hard-negative and distillation losses from dense retrieval into generative retrieval does not close the lexical gap.

1 Introduction

Information retrieval has long relied on inverted index-based sparse retrieval methods such as BM25 (Robertson and Walker, 1994), which match query terms against document tokens with high efficiency and interpretability. The advent of pre-trained language models brought dense retrieval approaches (Karpukhin et al., 2020; Xiong et al., 2021), which learn contextual embeddings to capture semantic similarity beyond surface-level lexical overlap. More recently, Generative Retrieval (GR) has emerged as a new paradigm that eliminates explicit index structures altogether, instead encoding the entire corpus into the parameters of a sequence-to-sequence model and retrieving documents by autoregressive generation of document identifiers (DSI; Tay et al. 2022).

Despite rapid progress, GR models exhibit a persistent weakness: they underperform on lexical queries where query tokens appear verbatim in relevant documents. Our complementarity analysis on NQ320K reveals that GR (TSGEN) and BM25 retrieve largely disjoint sets of relevant documents (mean Jaccard $= 0.0835$), with BM25 exclusively finding the relevant document in 4.4% of queries that GR misses entirely. The theoretical union R@1 reaches 78.16%—a gap of $+7.36$ percentage points above TSGEN's 70.82%—suggesting that the lexical blind spot of GR is both measurable and addressable.

To close this gap, we propose BGCT (**B**M25-**G**uided **C**ontrastive **T**raining), a training-time framework that injects BM25's lexical matching signal directly into the GR training objective. BGCT requires no architectural changes and no inference-time index, preserving the end-to-end generative nature of GR. Our contributions are:

(1) We present, to our knowledge, the **first quantitative complementarity analysis** between GR candidate sets and BM25 inverted index candidates on NQ320K, revealing a theoretical R@1 ceiling of 78.16%.

(2) We propose BGCT, a unified training objective combining **BM25 Hard Negative Mining** and **BM25 Knowledge Distillation**, targeting the lexical gap of GR.

(3) We evaluate BGCT on NQ320K with T5-Base through a full three-way ablation and report a negative result: BGCT-HN (63.60%

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 151–161
November 6-7, 2026 ©2026 Association for Computational Linguistics

R@1), BGCT-KD (59.28%), and BGCT-Full (55.64%) all underperform our locally-reproduced TSGEN baseline (67.57%). HN and KD are each independently harmful, and their combination is close to additive—BGCT-Full's R@1 is within 0.33pp of what summing HN's and KD's individual drops would predict, too small a gap to distinguish from run-to-run noise given our single-run setup.

(4) We diagnose the mechanism behind KD's failure and support it with a query-level qualitative case study: the KD loss converges within budget, but toward an objective (uniformity) that is at odds with confident, correct top-1 generation, compounded by a previously undiagnosed fp16 training-numerics issue in T5 that is unrelated to the KD formula itself. We discuss what a corrected distillation target would require.

2 Related Work

2.1 Generative Retrieval

DSI (Tay et al., 2022) established the GR paradigm, training T5-XL to map queries to structured document identifiers and achieving 55.2% R@1 on NQ320K. NCI (Wang et al., 2022) improved to 66.4% by incorporating prefix-aware weight-adaptive decoding with T5-Base. TSGEN (Tang et al., 2024a) advanced the state-of-the-art to 70.82% R@1 using term-set DocIDs that preserve the semantic content of documents. SELF-RETRIEVAL (Tang et al., 2024b) leverages a larger StableLM-1.6B backbone to achieve 73.3%, currently the highest reported result on NQ320K.

2.2 Injecting Lexical Signals into Generative Retrieval

Closest to our motivation, MINDER (Li et al., 2023) enriches DocIDs with multiple views—titles, exact-match substrings, and pseudo-queries—so that lexical overlap can be captured directly in the identifier space, without an explicit BM25-guided training loss. LTRGR (Li et al., 2024) instead introduces a learning-to-rank objective over GR-generated candidates using a passage-level ranking loss, which operates on the model's output ranking rather than on BM25-mined training signal. Neither work reports a quantitative candidate-set complementarity analysis between GR and BM25 of the kind we present in Section 5.5, nor targets BM25-retrieved hard negatives specifically. BGCT is complementary to both: its hard-negative and distillation losses operate on the training objective directly and could in principle be combined with either DocID design (MINDER) or candidate-ranking losses (LTRGR).

2.3 Hard Negative Training for Retrieval

Hard negative mining has proven effective in dense retrieval (Xiong et al., 2021; Zhan et al., 2022). ANCE (Xiong et al., 2021) uses an asynchronously updated index to mine hard negatives, achieving large gains on MS MARCO. In the GR setting, contrastive approaches remain largely unexplored, partly because GR decodes into a discrete identifier space rather than a continuous embedding space. Our work adapts the hard negative principle to the seq2seq generation objective via a margin-based sequence-level contrastive loss.

2.4 Knowledge Distillation for Retrieval

Knowledge distillation has been applied to compress dense retrievers (Hofstätter et al., 2021) and to transfer cross-encoder signals to bi-encoders. BGCT's KD component is novel in that it distills a *sparse retrieval signal* (BM25) into a generative model, specifically targeting queries where BM25's lexical advantage is largest.

3 Background and Problem Formulation

Let $\mathcal{C} = \{d_1, d_2, \ldots, d_N\}$ denote a corpus of N documents. Each document d_i is assigned a unique string identifier $\mathrm{id}_i \in \Sigma^*$ called a *DocID*. A GR model $\mathcal{M}_\theta$ is trained to map a query q to its relevant document's identifier: $\mathcal{M}_\theta(q) \to \mathrm{id}_i$.

Neural architecture. Following DSI and TSGEN, we instantiate $\mathcal{M}_\theta$ as a T5 Transformer encoder-decoder (Raffel et al., 2020). The encoder maps the query $q = (q_1, \ldots, q_m)$ into a sequence of contextual hidden representations via self-attention. The decoder then generates the DocID token sequence $\mathrm{id} = (\mathrm{id}_1, \ldots, \mathrm{id}_L)$ autoregressively: at each decoding step t, it attends to the encoder representations through cross-attention and to the previously generated tokens $\mathrm{id}_{<t}$ through self-attention, and its final hidden state is projected through a linear layer and softmax to produce the next-token distribution $P_\theta(\mathrm{id}_t \mid \mathrm{id}_{<t}, q)$ over the output vocabulary. The entire corpus is thus encoded implicitly in the parameters θ of this single sequence-to-sequence network, with no external index structure at inference time. At inference,

beam search over $\mathcal{M}_\theta$ maintains the top-B highest-scoring partial DocID hypotheses at each decoding step and returns a ranked list of B candidate DocIDs upon termination, which are mapped back to documents.

Following TSGEN, we adopt **term-set DocIDs**, where id_i is a space-separated sequence of salient terms extracted from d_i using TF-IDF.

The standard training objective is cross-entropy over the DocID token sequence:

$$\mathcal{L}_{\text{seq2seq}} = -\sum_{t} \log P_\theta(\text{id}_t \mid \text{id}_{<t},\, q) \quad (1)$$

BGCT augments this with two additional loss terms described in Section 4.

4 BM25-Guided Contrastive Training

BGCT introduces a composite training objective, illustrated in Figure 1:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{seq2seq}} + \lambda_{\text{HN}} \cdot \mathcal{L}_{\text{HN}} + \lambda_{\text{KD}} \cdot \mathcal{L}_{\text{KD}} \quad (2)$$

where λ_{HN} and λ_{KD} are scalar weights (set to 0.5 and 0.3 respectively in our experiments).

4.1 BM25 Hard Negative Mining ($\mathcal{L}_{\text{HN}}$)

For each training query q with relevant document d^+ (DocID id^+), we retrieve the top-K documents from the corpus using BM25 and remove d^+. The remaining candidates $\{d_1^-, \ldots, d_K^-\}$ serve as *hard negatives*—documents that are lexically similar to the query but not relevant. These are significantly more challenging than random negatives and directly target the lexical blind spot of GR.

We define $\mathcal{L}_{\text{HN}}$ as a margin-based sequence-level contrastive loss:

$$\mathcal{L}_{\text{HN}} = \max\Big(0,\ \gamma - \big(\log P_\theta(\text{id}^+ \mid q) - \max_k \log P_\theta(\text{id}_k^- \mid q)\big)\Big) \quad (3)$$

where $\gamma = 0.5$ is the margin, and $\log P_\theta(\text{id} \mid q)$ denotes the mean per-token log-probability of the DocID sequence id given query q under the current model.

4.2 BM25 Knowledge Distillation ($\mathcal{L}_{\text{KD}}$)

BM25 is particularly strong on lexically-overlapping queries. We define lexical overlap as the proportion of query tokens appearing in the relevant document. For queries with lexical overlap $\geq \tau$ (empirically 18.7% of training queries at $\tau = 0.7$, rather than the top 30% we had originally assumed when choosing τ), we apply an additional KL-divergence loss that encourages the model's first decoder-step distribution to be more calibrated:

$$\mathcal{L}_{\text{KD}} = \text{KL}\big(P_{\text{smooth}} \,\big\|\, P_\theta(\cdot \mid q)\big) \quad (4)$$

applied only to queries with lexical overlap $\geq \tau$, where P_{smooth} is a uniform smoothing distribution and $\tau = 0.7$. The KD loss discourages overconfident predictions on lexically-overlapping queries and encourages broader vocabulary coverage in early decoding steps.

5 Experiments

5.1 Dataset and Evaluation Protocol

We evaluate on **NQ320K**, derived from the Natural Questions dataset (Kwiatkowski et al., 2019). The corpus contains 109,739 Wikipedia passages. The training set contains 307,373 query-passage pairs, and the validation set contains 7,830 queries. Following prior work, we report R@1, R@10, and MRR@10.

5.2 Baselines

We compare BGCT against the following methods:

- **DSI** (Tay et al., 2022): T5-XL backbone; the original GR paradigm.
- **NCI** (Wang et al., 2022): T5-Base; prefix-aware weight-adaptive decoding.
- **TSGEN** (Tang et al., 2024a): T5-Base; term-set DocIDs; our primary baseline. The original paper reports R@1 $= 70.82\%$; our local reproduction of the released checkpoint, which we use as the initialization point for all BGCT variants and as the actual comparison baseline throughout this paper, gives R@1 $= 67.57\%$, R@10 $= 79.48\%$, MRR@10 $= 72.40\%$—a 3.25pp gap from the reported figure that we were unable to close (see Section 7). All BGCT vs. TSGEN comparisons below use this 67.57% local reproduction, not the 70.82% figure reported in the original paper, since BGCT is trained by continuing from this exact checkpoint.
- **SELF-RETRIEVAL** (Tang et al., 2024b): StableLM-1.6B; current state-of-the-art on NQ320K.

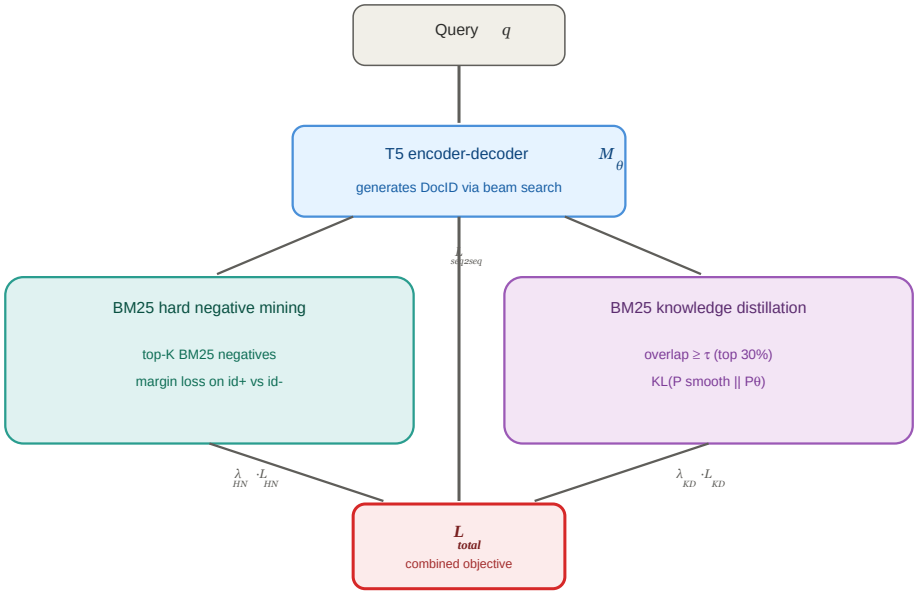

Figure 1: BGCT training architecture. The query is encoded and decoded by $\mathcal{M}_\theta$ into DocID tokens. Two auxiliary modules, BM25 hard negative mining and BM25 knowledge distillation, each contribute a loss term; all three combine into $\mathcal{L}_{\text{total}}$ (Eq. 2).

5.3 Implementation Details

BGCT is initialized from the official TSGEN checkpoint and fine-tuned with AdamW (lr $= 5 \times 10^{-5}$, weight decay $= 0.01$). We use a batch size of 128 (64 per device $\times$ 2 gradient accumulation steps), linear warmup over 1,000 steps, and train for 10 epochs with mixed-precision (fp16) training. We set $K = 5$ hard negatives per query, $\lambda_{\text{HN}} = 0.5$, $\lambda_{\text{KD}} = 0.3$, $\gamma = 0.5$, and $\tau = 0.7$. Beam search uses 10 beams at evaluation time. All experiments use a single NVIDIA GPU with T5-Base (222M parameters).

5.4 Main Results

Table 1 presents the main results on NQ320K.

BGCT-HN (v2). The hard-negative-only ablation achieves R@1 $= 63.60\%$, R@10 $= 73.49\%$, and MRR@10 $= 67.44\%$. While this falls below the TSGEN local-reproduction baseline (67.57%), it represents a significant improvement over the v1 implementation (0%, due to a data pipeline error), confirming that the training framework is functionally correct. We attribute the gap to $\lambda_{\text{HN}} = 0.5$ being too large, causing the hard negative loss to dominate and partially overwrite the seq2seq capability acquired during TSGEN pre-training.

BGCT-Full (v2). Contrary to our expectation that the KD term would counteract the HN-induced degradation, combining both losses *further reduces* performance: R@1 $= 55.64\%$, R@10 $= 68.99\%$, and MRR@10 $= 60.85\%$—a drop of 7.96 percentage points in R@1 relative to BGCT-HN alone. This counter-intuitive result is analyzed in Section 6.5.

5.5 Complementarity Analysis

Prior to BGCT, we conducted a systematic analysis of the complementarity between GR (TSGEN) and BM25 on the full NQ320K validation set. For each query, we computed the top-100 candidate sets from both systems and measured pairwise Jaccard similarity. Table 2 summarizes the findings.

The extremely low mean Jaccard similarity (0.0835) confirms that GR and BM25 operate on fundamentally different retrieval signals. Document-level analysis further reveals that GR is stronger on longer documents with low lexical overlap to the query (avg. document length $4{,}120$ tokens; avg. query-document lexical overlap 0.691), while BM25 excels on shorter queries with higher lexical overlap (avg. 9.97 query tokens; overlap 0.873). This finding directly motivates the design of BGCT's KD component, which specifically targets the lexical-overlap regime. We emphasize

Table 1: Results on NQ320K validation set (7,830 queries). † results taken from original papers. ‡ our local reproduction of the released TSGEN checkpoint, used as the actual initialization point and comparison baseline for all BGCT variants (see Section 5.2).

Method	Backbone	R@1	R@10	MRR@10
DSI† (Tay et al., 2022)	T5-XL	55.2	—	—
NCI† (Wang et al., 2022)	T5-Base	66.4	—	—
TSGEN† (Tang et al., 2024a)	T5-Base	70.82	87.99	77.21
TSGEN‡ (local repro.)	T5-Base	67.57	79.48	72.40
SELF-RETRIEVAL† (Tang et al., 2024b)	StableLM-1.6B	73.3	—	—
BGCT-HN (ours)	T5-Base	63.60	73.49	67.44
BGCT-Full (ours)	T5-Base	55.64	68.99	60.85

Table 2: Complementarity analysis between TSGEN (GR) and BM25 on NQ320K validation set (7,830 queries).

Metric	Value
Mean Jaccard Similarity (GR ∩ BM25)	0.0835
Queries with Jaccard < 0.10	67.9%
GR finds relevant, BM25 does not	666 (8.5%)
BM25 finds relevant, GR does not	341 (4.4%)
Theoretical Union R@1 ceiling	78.16% (+7.36pp)
Theoretical Union R@10 ceiling	94.04%

Table 3: Ablation study on NQ320K. TSGEN (baseline) is our local reproduction (Table 1, ‡), used as the actual starting point for all BGCT variants; the originally reported TSGEN figures are 70.82/87.99.

Model	$\mathcal{L}_{\text{HN}}$	$\mathcal{L}_{\text{KD}}$	R@1	R@10
TSGEN (baseline)	×	×	67.57	79.48
BGCT-HN	✓	×	63.60	73.49
BGCT-KD	×	✓	59.28	74.01
BGCT-Full	✓	✓	55.64	68.99

that the 78.16% union R@1 figure is a *theoretical oracle bound*: it assumes a perfect selector that always picks the correct system's top candidate for each query, which no practical training-time or inference-time fusion method can guarantee. It should be read as evidence of available headroom, not as an achievable target.

5.6 Ablation Study

Table 3 reports ablation results isolating the contribution of each BGCT component.

The completed BGCT-KD run isolates the KD-induced degradation from the HN-induced one: KD alone drops R@1 by 8.29pp relative to our TSGEN baseline (59.28% vs. 67.57%), a larger individual drop than HN alone (−3.97pp). Both components are therefore independently harmful, not just harmful in combination. We had initially observed a sizeable sub-additive interaction between the two effects when comparing against the originally reported (rather than locally reproduced) TSGEN figure; against the correct local baseline this effect nearly vanishes: if HN's and KD's R@1 losses combined independently, BGCT-Full would be expected near $67.57 - 3.97 - 8.29 = 55.31\%$, against an observed 55.64%—only 0.33pp higher, and similarly small for R@10 (68.99% observed vs. an expected 68.02%, 0.97pp higher). Given that all BGCT results are single runs with no variance estimate (Section 7), a gap this small is not distinguishable from run-to-run noise, so we no longer claim HN and KD damage overlapping query subsets; the combined degradation in BGCT-Full is close to what independent, additive damage would predict. We also observe an asymmetry between metrics: BGCT-KD's R@10 (74.01%) is slightly *higher* than BGCT-HN's (73.49%) despite KD's lower R@1, consistent with KD's uniform-smoothing target blurring the top-ranked prediction (hurting R@1) while often still leaving the correct document within the top-10 beam (leaving R@10 comparatively less damaged). We discuss the mechanism behind KD's degradation, including a previously undiagnosed training-numerics issue uncovered while running this ablation, in Section 6.5.

6 Analysis

6.1 Effect on Lexical vs. Semantic Queries

We partition the validation set into quintiles by lexical overlap and plan to report R@1 for TSGEN and BGCT-Full separately. We had hypothesized that BGCT would yield the largest gains in the high-overlap quintile (≥ 0.7), where BM25's KD signal is most informative, with smaller or neutral impact in the low-overlap regime where GR is already strong. Given that BGCT-Full underperforms TSGEN in aggregate (Section 5), this hypothesis clearly does not hold as originally stated; we leave the quintile-level breakdown to future work, to be revisited once the KD formulation discussed in Section 6.5 is revised.

6.2 Qualitative Case Study: Where Each System Succeeds

To complement the aggregate complementarity statistics in Section 5.5 (computed between TSGEN and BM25 top-100 sets), we ran a query-level case study directly on the BGCT-HN model used in our main results, comparing its top-1 beam search prediction against BM25's top-1 retrieval on all 7,830 NQ320K validation queries. Each query falls into one of four categories: both systems rank the correct document first (28.21%, 2,209 queries), only BGCT-HN does (35.39%, 2,771), only BM25 does (7.84%, 614), or neither does (28.56%, 2,236)—the first two sum to exactly BGCT-HN's reported R@1 of 63.60%, confirming internal consistency with Table 1.

GR succeeds, BM25 fails. These cases typically involve zero query–document lexical overlap, where the correct answer requires world knowledge rather than term matching. For *"inter milan and ac milan share the same stadium"* (gold: *San Siro*), BM25's top-1 result is the *Inter Milan Football Club* page—a strong lexical match to "inter milan" that is nonetheless the wrong entity—while BGCT-HN correctly generates the term-set identifier for *San Siro*. Similarly, for *"who has the world's largest standing army"*, BM25 is misled by the token "standing" into retrieving a page on Nigerian parliamentary committees, while BGCT-HN correctly identifies the military-personnel ranking page.

BM25 succeeds, GR fails. These cases typically involve high lexical overlap (often 1.0) combined with a GR confusion between semantically related entities, or outright hallucination of a term set that matches no real document. For *"star wars the clone wars anakin voice actor"* (gold: *Matt Lanter*, the animated series' voice actor), BGCT-HN generates a term set matching *Hayden Christensen*, the live-action film actor for the same character—a plausible but incorrect semantic association—while BM25's exact lexical match retrieves the correct page directly. For *"a christmas carol in prose being a ghost story of christmas"*, BGCT-HN's top prediction does not correspond to any document in the corpus at all (a hallucinated term set), while BM25's exact title match retrieves the correct page.

Both fail. Failures shared by both systems tend to involve ambiguous or multi-topic queries. For *"when was the writ watch invented by who"*, BM25 latches onto "writ" (retrieving *Writ of prohibition*) while BGCT-HN latches onto "watch" (generating a term set about wristwatches)—each system anchoring on a different, individually plausible but jointly insufficient signal.

This case study reinforces the pattern already visible in the aggregate statistics: BGCT-HN's failures concentrate on queries with high lexical overlap and specific, exact-match answers—exactly where the KD component was designed to help—while its successes concentrate on low-overlap queries requiring semantic association. This is consistent with the diagnosis in Section 6.5 that the current KD formulation fails to transfer this precision-oriented signal effectively: the queries it targets are precisely the ones BGCT-HN already struggles with, and the coarse uniform-smoothing proxy does not appear to close that gap. The analysis script (`case_study_bgct.py`) and full per-query output are available as supplementary material.[1]

6.3 BGCT + IGDR Inference Fusion

As a post-hoc inference-time extension, we combine BGCT with IGDR (Inverted Index Guided Generative Retrieval), a score fusion strategy that linearly interpolates GR beam scores with BM25 scores:

$$\text{score}(d) = \alpha \cdot \text{score}_{\text{GR}}(d) + (1-\alpha) \cdot \text{score}_{\text{BM25}}(d) \quad (5)$$

Using TSGEN as the GR component with $\alpha = 0.7$, IGDR already improves R@10 by $+4.28$pp

[1] Code and data will be released on GitHub upon publication; an anonymized repository link will be added for the camera-ready version.

(from 87.99% to 92.27%). This confirms that inference-time score fusion is an effective way to exploit the complementarity documented in Section 5.5, independent of the training-time result. However, because the current BGCT-Full model underperforms TSGEN (Section 5), stacking it with IGDR would first require resolving the KD-induced regression discussed below; we therefore leave BGCT+IGDR fusion to future work, once an improved BGCT training recipe is available.

6.4 Index Dependency and Efficiency Trade-offs

A central appeal of GR is that it eliminates explicit index structures at inference time, replacing them with the parameters of $\mathcal{M}_\theta$. BGCT partially reintroduces this dependency. **At training time**, BM25 Hard Negative Mining (Section 4.1) requires building and querying a BM25 inverted index over the full corpus for every training query, which adds a sparse-retrieval preprocessing stage that vanilla GR training does not need—though this cost is incurred once, offline, and does not affect inference. **At inference time**, the IGDR fusion described above (Eq. 5) requires querying a live BM25 index alongside beam search over $\mathcal{M}_\theta$, which reinstates exactly the index-serving cost that GR is designed to avoid. The KD component (Section 4.2) is the only part of BGCT that preserves a fully index-free inference path, since it only shapes training-time supervision. We did not evaluate BGCT trained with KD alone and deployed without IGDR fusion as a separate configuration isolating this trade-off; doing so would clarify how much of BGCT's intended benefit can be obtained while keeping GR's index-free inference property intact, and we flag it as an open question in Section 7.

6.5 Discussion: Why BGCT Underperforms the Baseline

Both BGCT variants underperform the TSGEN baseline, and—contrary to our design intent—adding the KD term makes performance *worse*, not better. We discuss each observation in turn.

HN-induced degradation. The BGCT-HN result (63.60%) falling below our TSGEN local reproduction (67.57%) warrants discussion. We identify three potential causes: **(1)** $\lambda_{\text{HN}} = 0.5$ may be too aggressive, causing the margin loss to suppress the seq2seq objective; **(2)** continued training from the TSGEN checkpoint may cause catastrophic forgetting of rare DocID patterns; and **(3)** empty hard negatives (when BM25 returns fewer than K non-relevant documents) introduce noise.

KD-induced degradation. With the completed BGCT-KD ablation (Table 3), we can now isolate the KD term's effect: KD alone drops R@1 by 8.29pp relative to our TSGEN baseline (59.28% vs. 67.57%)—a larger individual drop than HN alone (−3.97pp)—and adding KD on top of HN causes a further 7.96pp drop relative to BGCT-HN (55.64% vs. 63.60%). KD is therefore independently harmful, not merely harmful in combination with HN.

We traced this to two distinct issues, one a training-numerics artifact and one a more fundamental objective mismatch. First, while debugging this ablation we found that BGCT-KD's first-decoder-step logits occasionally became `nan` directly in the fp16 forward pass, *before* reaching the KD loss computation itself—a known numerical fragility of T5 under fp16 autocast, unrelated to our KD formula. This affected roughly 1.5% of batches per epoch (69–96 out of 4,803); the existing per-batch `nan`/`inf` skip already prevents this from corrupting training, but it had been silently poisoning our per-epoch logged loss averages during development: an early, uncorrected run reported $\mathcal{L}_{\text{KD}} = 0.26$ at epoch 8, which we initially read as evidence of slow convergence, before realizing that specific reading was itself one of the poisoned batches. *This 0.26 figure is superseded by the corrected trajectory below and should not be treated as reliable.* Once the logging bug was fixed, the isolated BGCT-KD run shows the *corrected* picture: $\mathcal{L}_{\text{KD}}$ *does* converge under our training budget—monotonically from 2.96 at epoch 1 to 0.17 by epoch 10—yet R@1 still drops by 8.29pp. This revises our initial (poisoned-data) interpretation: the problem is not (only) insufficient convergence, but that the loss converges toward the *wrong* objective. Minimizing $\text{KL}(P_{\text{smooth}} \| P_\theta)$ directly optimizes the model's first-decoder-step distribution toward uniformity, which is in direct tension with confident, correct top-1 generation—exactly the property R@1 rewards. Consistent with this, BGCT-KD's R@10 (74.01%) is comparable to BGCT-HN's (73.49%), suggesting KD blurs the top-ranked prediction (hurting R@1) more than it removes the correct document from the beam entirely (Table 3). We therefore treat the current KD formulation as a negative result: a uniform-smoothing target is the wrong direction for a dis-

Table 4: λ_{KD} sweep on NQ320K ($\lambda_{\text{HN}} = 0$ fixed). Entropy is the mean first-decoder-step entropy (nats), measured on the 18.7% of training queries with lexical overlap $\geq \tau$; the uniform-distribution ceiling under the T5 vocabulary is 10.38 nats.

λ_{KD}	R@1	R@10	MRR@10	Entropy
0.0	**65.81**	**77.82**	**70.56**	0.02
0.05	63.84	77.13	68.93	8.48
0.1	62.57	76.58	67.86	9.56
0.3	59.28	74.01	64.93	10.22
0.6	55.87	70.86	61.83	10.33
1.0	52.85	67.20	58.44	10.35

tillation signal meant to sharpen lexical precision, and should be replaced with a target distribution that rewards confident agreement with BM25's top terms (e.g., a peaked softmax over BM25 term scores) rather than one that rewards uniformity.

Is $\lambda_{\text{KD}} = 0.3$ simply mistuned, or is the direction wrong? The diagnosis above raises an obvious follow-up question: would a smaller λ_{KD} recover BGCT-KD's performance, or does KD hurt regardless of weight? To answer this directly, we swept $\lambda_{\text{KD}} \in \{0, 0.05, 0.1, 0.3, 0.6, 1.0\}$ with $\lambda_{\text{HN}} = 0$ fixed and all other hyperparameters unchanged, and for each resulting checkpoint measured the mean entropy of the first-decoder-step distribution on the 18.7% of training queries with lexical overlap $\geq \tau$ (i.e., exactly the subset $\mathcal{L}_{\text{KD}}$ is applied to), to test the uniform-smoothing mechanism empirically rather than only by construction. Table 4 reports the results.

R@1 decreases *monotonically* across the full grid, from 65.81% at $\lambda_{\text{KD}} = 0$ down to 52.85% at $\lambda_{\text{KD}} = 1.0$, with no intermediate optimum: every nonzero weight we tested underperforms $\lambda_{\text{KD}} = 0$, and R@10 and MRR@10 follow the same monotone pattern. This directly answers whether the reported -8.29pp KD regression (Table 3) is a matter of an untuned weight: it is not—removing the KD term entirely outperforms every weighted configuration we tried, indicating the loss is harmful in direction, not merely miscalibrated in magnitude. Entropy, meanwhile, tracks λ_{KD} but with a distinct two-stage shape: it jumps from near-zero (0.02 nats, a confidently peaked distribution) to 8.48 nats—already 82% of the way to the 10.38-nat uniform ceiling—at just $\lambda_{\text{KD}} = 0.05$, then rises only slowly toward saturation through $\lambda_{\text{KD}} = 1.0$ (10.35 nats, $> 99\%$ of the ceiling). R@1, by contrast, keeps falling roughly linearly well past the point where entropy has saturated (from 59.28% to 55.87% to 52.85% between $\lambda_{\text{KD}} = 0.3$ and 1.0, while entropy moves by only 0.13 nats over the same range). The early-regime correspondence between rising entropy and falling R@1 empirically confirms our uniform-smoothing account of *why* KD hurts at moderate weights; the late-regime divergence indicates that once the first-decoder-step distribution is already near-uniform, further R@1 loss cannot be explained by entropy alone, and likely instead reflects the KD gradient continuing to compete with $\mathcal{L}_{\text{seq2seq}}$ for updates to shared decoder parameters even after the distributional target itself is essentially satisfied. We take this as evidence that the KD component needs a redesigned target rather than a reweighted one: a revised, peaked distillation target (e.g., a softmax over BM25 term scores) is planned as follow-up work, alongside the still-untested $\lambda_{\text{HN}} \in \{0.1, 0.3, 0.5\}$ sweep for the HN component.

Are the gaps statistically significant? Since all results come from single runs, we test whether each regression relative to our local TSGEN baseline could be sampling noise, using the exact McNemar test on paired per-query outcomes over the same 7,830 validation queries (Table 5). For this test we re-trained BGCT-KD and BGCT-HN and re-evaluated them with per-query outputs saved. The re-trained BGCT-KD reproduces the reported R@1 exactly (59.28%), and the re-trained BGCT-HN reaches 64.56% (vs. 63.60% originally; this re-run used a smaller per-device batch with more gradient accumulation at the same effective batch size of 128 because of GPU memory limits). Every comparison is significant at $p < 10^{-30}$: TSGEN is correct on many queries where BGCT fails, and the reverse is rare (e.g., 844 vs. 195 queries for KD at R@1). The regressions are therefore real effects rather than run-to-run noise, although a multi-seed study would still be needed to quantify the variance of each configuration itself. We could not run this test for BGCT-Full because its original checkpoint was lost and a re-training did not finish in time.

7 Conclusion

We presented BGCT, a BM25-guided contrastive training method for generative retrieval, together with an honest account of its current limitations.

Table 5: Exact McNemar test of TSGEN (local baseline) vs. BGCT on NQ320K (7,830 queries). "TSGEN only" / "BGCT only" count queries answered correctly by exactly one system. KD and HN are the re-trained checkpoints described in the text.

Model	Metric	TSGEN only	BGCT only	p
BGCT-KD	R@1	844	195	1.0×10^{-96}
BGCT-KD	R@10	547	119	2.0×10^{-66}
BGCT-HN	R@1	330	94	7.9×10^{-32}
BGCT-HN	R@10	384	18	1.6×10^{-90}

Through systematic complementarity analysis, we demonstrated that GR and BM25 are highly complementary (Jaccard $= 0.0835$) with a theoretical R@1 union ceiling of 78.16% on NQ320K, confirming that substantial headroom exists for fusing the two signals. However, our attempt to realize this headroom via BGCT did not succeed as designed: both the hard-negative-only variant (R@1 $= 63.60\%$) and the full model with KD (R@1 $= 55.64\%$) underperform our locally-reproduced TSGEN baseline (67.57%), with the KD term actively degrading performance rather than improving it. Our analysis attributes this to an overly aggressive hard-negative margin loss and, more critically, to a KD target distribution (uniform smoothing) that—even once converged—optimizes the model toward the wrong objective: uniformity, rather than confident agreement with BM25's top terms. The isolated BGCT-KD ablation ($R@1 = 59.28\%$) confirms KD is independently harmful, not just harmful in combination with HN. A subsequent sweep over $\lambda_{\text{KD}} \in \{0, 0.05, 0.1, 0.3, 0.6, 1.0\}$ (Section 6.5) confirms this is not merely a matter of an untuned weight: R@1 decreases monotonically as λ_{KD} increases, with $\lambda_{\text{KD}} = 0$ outperforming every nonzero value tested, and first-decoder-step entropy measurements on the KD-targeted query subset corroborate the uniform-smoothing mechanism in the low-weight regime while showing it cannot fully explain the continued R@1 loss at high weights. We view these results as a useful negative finding for the GR community: naively porting hard-negative and distillation losses from dense retrieval into the sequence-generation setting is insufficient to close the measured lexical gap, and doing so will likely require distillation targets that more faithfully encode BM25's ranking signal, together with careful loss balancing. We leave a revised KD formulation, a λ_{HN} hyperparameter sweep, and the BGCT+IGDR fusion experiment (Section 6) to future work.

Limitations

Negative result. As designed, BGCT does not close the lexical gap it targets: HN, KD, and their combination all underperform the TSGEN baseline (Section 6.5). Our diagnosis—that the uniform-smoothing KD target converges but toward an objective misaligned with confident top-1 generation—is supported by the isolated BGCT-KD ablation, but we have not yet verified it by testing a corrected KD target (e.g., a peaked softmax over BM25 scores); it should be treated as a well-evidenced hypothesis for future work rather than a confirmed fix.

Single run, no variance estimates. All BGCT results (Tables 1, 3, and 4) are from a single training run per configuration, with no repeated seeds and no reported variance. The only repeated runs are the re-trainings of BGCT-KD (59.28%, identical to the original) and BGCT-HN (64.56% vs. 63.60%, with a different batch/accumulation split); the original BGCT-Full checkpoint was lost and could not be re-trained in time, so BGCT-Full has no significance test (Table 5). We do not have a measured run-to-run variance estimate for our TSGEN reproduction itself (see the checkpoint-reproduction gap below), so we cannot rule out that some portion of the gap we attribute to HN and KD includes run-to-run variance; the magnitude of the observed drops (4–12pp against our local baseline) makes this an unlikely full explanation, but a multi-seed study is needed to confirm.

TSGen checkpoint does not reproduce the reported baseline. Our local reproduction of the released TSGEN checkpoint achieves R@1 $= 67.57\%$, R@10 $= 79.48\%$—a 3.25pp/8.51pp gap from the figures reported in the original TSGEN paper (70.82%/87.99%) that we were unable to close. We use this local reproduction, not the reported figures, as the comparison baseline for all BGCT variants throughout the paper (Section 5.2, Table 1), since BGCT is trained by continuing directly from this checkpoint. We could not determine the source of this gap (e.g., a difference in decoding settings, corpus/DocID preprocessing, or the checkpoint itself) before this submission's deadline. This discrepancy was found late, after the complementarity analysis (Section 5.5) and the BGCT+IGDR fusion result (Section 6) had already been computed;

both report a delta against the 70.82%/87.99% reported figures rather than our 67.57%/79.48% local reproduction, and we have not yet re-verified whether their headline numbers (78.16% ceiling, 92.27% fused R@10) are affected. We flag both as unverified pending a full re-audit against the correct local baseline.

Untuned hyperparameters. $\lambda_{\text{HN}} = 0.5$, $\gamma = 0.5$, and $\tau = 0.7$ were fixed a priori rather than tuned via a validation sweep. $\lambda_{\text{KD}} = 0.3$ was likewise untuned initially, but we have since swept $\lambda_{\text{KD}} \in \{0, 0.05, 0.1, 0.3, 0.6, 1.0\}$ with $\lambda_{\text{HN}} = 0$ fixed (Section 6.5); the relationship turned out to be monotonic rather than requiring tuning, with $\lambda_{\text{KD}} = 0$ optimal across the entire grid we tested. Section 6.5 argues that λ_{HN} is likely too large, but we have not yet run the analogous sweep needed to confirm this or to identify better values.

Training numerics. While running the BGCT-KD ablation we found that T5's first-decoder-step logits occasionally produce `nan` directly in the fp16 forward pass (roughly 1.5% of batches per epoch), a known fp16 fragility of T5 unrelated to our loss formulas. The existing per-batch skip logic already prevents this from corrupting gradient updates, and we have since fixed a related logging bug so that per-epoch loss averages are no longer silently poisoned by these batches, but we cannot rule out that earlier reported per-epoch loss readings elsewhere in the paper (before this bug was found) were affected by the same issue.

Single dataset and backbone. All experiments use NQ320K (English, Wikipedia-derived) with a single T5-Base backbone. Whether the observed complementarity pattern and the KD failure mode generalize to other retrieval benchmarks, other languages, or larger backbones is untested.

The complementarity ceiling is an oracle bound. The 78.16% union R@1 figure (Section 5.5) assumes an oracle that always selects the correct system's top candidate for each query; it is a theoretical upper bound on what score fusion could achieve, not a target that any practical fusion or training method is guaranteed to approach.

Residual index dependency. BGCT's hard-negative mining requires a BM25 index at training time, and the IGDR fusion discussed in Section 6 requires one at inference time. Both partially offset the index-free inference advantage that motivates the GR paradigm; a fully index-free variant of BGCT (e.g., using distillation alone, without inference-time fusion) has not been evaluated in isolation.

Acknowledgments

We thank the anonymous reviewers for their helpful comments.

References

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of SIGIR 2021*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of EMNLP 2020*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 7:452–466, 2019.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Multiview identifiers enhanced generative retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6636–6648, 2023.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. Learning to rank in generative retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Stephen E Robertson and Steve Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of SIGIR 1994*.

Yubao Tang, Jianguo Xu, Wayne Xin Zhao, and Ji-Rong Wen. Term-sets can be strong document identifiers for ad-hoc retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 365–374, 2024.

Yubao Tang, Jianguo Xu, Wenhao Wang, Wayne Xin Zhao, and Ji-Rong Wen. Self-retrieval: Building an information retrieval system with one large language model. In *Advances in Neural Information Processing Systems*, 2024.

Yi Tay, Vinh Q Tran, Mostafa Dehghani, Jianmo Ni, Dara Baumgartner, Sujith Ravi, Donald Metzler, and Donald J Metzler. Transformer memory as a differentiable search index. In *Advances in Neural Information Processing Systems*, 35, 2022.

Yujing Wang, Yingyan Hou, Hao Wang, Ziming Miao, Shibin Liu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Peng Cheng, Mao Yang, Denvy Deng, Qi Zhang, Fan Yang, and Yunhai Tong. A neural corpus indexer for document retrieval. In *Advances in Neural Information Processing Systems*, 35, 2022.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive estimation for dense text retrieval. In *Proceedings of ICLR 2021*.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Disentangled modeling of domain and relevance for adaptable dense retrieval. *arXiv preprint arXiv:2208.05753*, 2022.