DRIFT-LOCALIZED ORDER EFFECTS IN TEMPORAL DENSE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Temporal retriever refresh must reconcile examples that add facts with later passages that revise or contradict them, but ordinary curriculum comparisons entangle their order with exposure, replay, and negative mining. We identify the residual sequence effect by materializing a 6.4-million-occurrence multiset containing primary and 20% replay triplets before training and permuting only their occurrence-to-step assignment; every occurrence retains its positive and hard negative, and the loss has no cross-example terms. A frozen six-signal drift card defines response strata, while STRATAORDER schedules additive, revision, and conflict phases forward. On Reuters News, versioned API documentation, and Wiki Entities, Contriever-MSMARCO reaches 74.7 Recall@100 in the eight-window high-drift slice versus 70.1 for global random order ($+4.6$, 95% CI $[3.1, 6.2]$, $d = 0.89$) and 72.4 for random order with explicit time tokens ($+2.3$, $[0.7, 3.8]$). Reverse ordering jointly reduces final-state recall and current-over-stale preference and increases forgetting; phase-preserving shuffling lies between Random and full forward order. ColBERTv2 reproduces the high-drift contrast (+4.0 [2.5, 5.5]), while both architectures show only $+0.3$ in the predeclared null slice. The resulting drift–order map turns sequence design from a generic curriculum choice into a measurable property of retrospective refresh: revisions and conflicts identify where fixed-exposure scheduling materially changes retrieval.

1 INTRODUCTION

An API symbol can be introduced in one release, revised in the next, and assigned an incompatible behavior in a third. A retriever refreshed retrospectively on all three states can see exactly the same training triplets yet rank different versions at the end, depending on which state occupied each optimization step. The consequence is specific to temporal retrieval: a passage may match the topic while reporting last quarter's event state, obsolete API behavior, or a superseded entity value.

Dense retrieval research has improved representation quality, negative mining, and late interaction (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021; Khattab & Zaharia, 2020; Santhanam et al., 2022; Izacard et al., 2022), while temporal benchmarks test access to changing knowledge (Liska et al., 2022; Kasai et al., 2023; Jang et al., 2022; Vu et al., 2024). Neither line resolves a basic attribution problem in refresh training. Changing a schedule commonly also changes replay frequency, example multiplicity, negative assignments, batch objectives, or update budget, so a better final retriever cannot be attributed to order itself.

Retrospective refresh makes the sequence independently manipulable because all corpus versions are visible before optimization. Our matched-multiset protocol first instantiates every query–positive–negative occurrence needed for 50K steps, including 20% replay, and only then assigns occurrences to steps. The mini-batch objective averages independent triplet losses, eliminating in-batch, cross-device, and queue negatives as hidden schedule-dependent inputs. A frozen six-component drift card classifies the evaluated windows before test scoring. STRATAORDER then presents additive windows, revisions, and direct conflicts in that order while preserving chronology within each phase; Random, Reverse, and within-phase shuffling probe distinct disruptions of the same occurrence multiset.

The technical contribution is an identification-and-localization protocol, not an average curriculum comparison. Across Reuters News, API Docs, and Wiki Entities, its high-drift contrast is $+4.6$ Recall@100 over Random ($[3.1, 6.2]$) and $+2.3$ over a random-order retriever with explicit time tokens

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

([0.7, 3.8]); the Random contrast contracts to +1.8 [0.7,3.0] at medium drift and $+0.3[-0.4, 1.1]$ in the test-frozen null slice. Current-over-stale preference and retention move with recall under destructive reorderings, and the high/null boundary persists under alternative drift definitions, 25K–75K budgets, and ColBERTv2. Together, the matched intervention, directional controls, and response map establish a practical conclusion for the evaluated retrospective streams: schedule choice is consequential precisely where corpus states revise or contradict one another.

2 Related Work

Temporal retrieval and model updates. Static retrieval benchmarks (Kwiatkowski et al., 2019; Joshi et al., 2017; Thakur et al., 2021) and temporal resources (Liska et al., 2022; Kasai et al., 2023; Jang et al., 2022; Vu et al., 2024) measure complementary capabilities. DynamicIR compares indexing- and training-based corpus updates (Kim et al., 2024), and GradNormIR predicts when a shifted corpus warrants retriever adaptation (Ko et al., 2025). Replay and regularization address retention (Kirkpatrick et al., 2017; Lopez-Paz & Ranzato, 2017; Rebuffi et al., 2017; Chaudhry et al., 2019; Aljundi et al., 2019; Rolnick et al., 2019; Buzzega et al., 2020; Sun et al., 2022); L2R changes representation compatibility (Cai et al., 2023), DSI++ updates a generative index (Mehta et al., 2023), Query Drift Compensation aligns updated queries with existing document embeddings (Goswami et al., 2026), and CREAM learns memory for an unlabeled stream (Son et al., 2026). These works decide whether or how to update. Our estimand instead holds the refresh data, negatives, replay, objective, and budget fixed and measures the remaining sensitivity to their sequence.

Curriculum and order interventions. Curriculum learning arranges examples by difficulty or competence (Bengio et al., 2009; Kumar et al., 2010; Jiang et al., 2018; Platanios et al., 2019; Hacohen & Weinshall, 2019); retrieval variants order distillation difficulty or pseudo-query relevance (Zeng et al., 2022; He et al., 2023), and fixed-data reasoning experiments show interactions among model, task, and direction (Jia et al., 2026). Temporal retrieval supplies a different ordering variable: the semantic relation between successive corpus states. By fixing every training occurrence and pairing a six-signal drift card with forward, reverse, global-random, and phase-preserving permutations, this work identifies where version semantics make sequence consequential and tests the same response with freshness, retention, and cross-architecture evidence.

3 Matched-Multiset Identification of Order Effects

3.1 Setting and estimand

Let $W_1, \ldots, W_T$ be versioned corpus windows. Each W_t contributes supervised query–positive pairs P_t and a temporally aligned evaluation set Q_t. All windows are observable before training, and evaluation queries remain held out. Before optimization, each training pair is joined to one fixed hard negative, producing a triplet occurrence $o = (q, p^+, p^-)$. A training sequence π orders the complete multiset $\mathcal{M}$ of primary and replay triplet occurrences. For metric Y and drift slice G, the residual order effect is

$$\Delta_{\mathrm{ord}}(G; \pi_1, \pi_0) = \mathbb{E}[Y(\pi_1) - Y(\pi_0) \mid W_t \in G,\, \mathcal{M}_{\pi_1} = \mathcal{M}_{\pi_0}]\,. \tag{1}$$

Equality covers the identity and multiplicity of every presented triplet and its primary/replay designation. The learning-rate schedule is the same function of optimizer step for every method, but the occurrence mapped to a step changes with π; fixing that mapping would eliminate the intervention. Paired runs additionally share initialization, batch count, and total updates. Under these conditions, Δ_{ord} isolates the contrast between presented sequences. Evaluating that contrast over strata frozen before test scoring produces the drift–order response map

$$\mathcal{R}(\pi_1, \pi_0) = \{\Delta_{\mathrm{ord}}(G; \pi_1, \pi_0) : G \in \{\mathrm{high}, \mathrm{medium}, \mathrm{null}\}\}, \tag{2}$$

which preserves heterogeneity that a stream-wide average would merge.

Table 1 separates matched-order controls from practical baselines. Random globally permutes $\mathcal{M}$. Reverse orders conflict, revision, then additive phases and reverses window order inside each phase. Shuffled-Within-Phase preserves phase membership but independently permutes occurrences inside every phase. StrataOrder uses forward phases and chronological within-phase order.

Table 1: Comparator schema. "Same multiset" denotes equality of all query–positive–negative occurrences and repetition counts with the matched-order family.

Method	Replay	Same multiset	Time input	Extra params.	Role
Random	20%	yes	no	0	global shuffle
Reverse	20%	yes	no	0	direction
Shuffled-Within-Phase	20%	yes	no	0	local order
StrataOrder	20%	yes	no	0	forward strata
Time-Conditioned	20%	yes	yes	12,288	random order + time input
Replay-only	20%	no	no	0	replay baseline
EWC-only	0	no	no	0	regularization
Replay+EWC	20%	no	no	0	tuned hybrid
Latest-Only	0	no	no	0	latest-window

Time-Conditioned uses Random's permutation and adds release tokens, making Random–Time-Conditioned the token contrast and StrataOrder–Time-Conditioned a comparison between two temporal design choices.

Sequence construction. For each stream and seed, sequence generation is completed before any model update:

1. Split assignment and pair construction produce the eligible primary pool. For each pair, the hard negative is the highest-scoring eligible stale or superseded training passage under the frozen initialization. Phase assignment is completed before occurrence instantiation.
2. A deterministic balanced sampler cycles over primary identifiers to fill exactly $50{,}000 \times 128 = 6.4$ million slots. After every four primary IDs, seeded round-robin sampling selects one replay ID from the phase-eligible source pool. This fixes 20% of slots as replay before ordering; the resulting occurrence list, rather than the unique-pair set in Table 2, is $\mathcal{M}$.
3. An ordering key maps the same occurrence IDs to positions: a seeded global permutation for Random and Time-Conditioned; phase then time then seeded tie-break for StrataOrder; phase then an independent within-phase permutation for Shuffled-Within-Phase; and reversed phase and time keys for Reverse. Because the study is retrospective, replay is an exposure designation rather than an online availability claim; an occurrence may move before its source phase under a destructive permutation.
4. Consecutive blocks of 128 occurrences form batches without padding. Optimizer step k receives the same learning-rate value and update rule across paired methods, while π determines which block occupies k.

Thus all matched-order methods share 6.4 million occurrence IDs, exactly 1.28 million replay-designated presentations, hard-negative assignments, and per-pair multiplicities.

Order-only objective. Every occurrence uses a two-passage softmax, $\ell_o = \log(1+\exp[s_\theta(q,p^-) - s_\theta(q,p^+)])$, where s_θ is a single-vector inner product for Contriever-MSMARCO and token-level MaxSim for ColBERTv2. The step loss averages 128 independent occurrence losses. Other batch members never enter an occurrence's denominator, and training uses no in-batch, cross-device, queued, or newly mined negatives. Permutation therefore changes when fixed triplets contribute gradients without changing any per-occurrence objective.

3.2 Frozen drift card

For each adjacent transition, we compute a six-component descriptor from training data:

$$\mathbf{d}_t = [\mathrm{LexShift}_t, \mathrm{EmbDisp}_t, \mathrm{AnsFlip}_t, \mathrm{EntTurn}_t, \mathrm{ContrDens}_t, \mathrm{QDMismatch}_t]. \quad (3)$$

Lexical shift is one minus Jaccard overlap between lowercased unigram vocabularies in adjacent windows. Embedding displacement is the cosine distance between centroids of mean-pooled, L2-normalized passage embeddings from the untouched Contriever-MSMARCO checkpoint. Answer

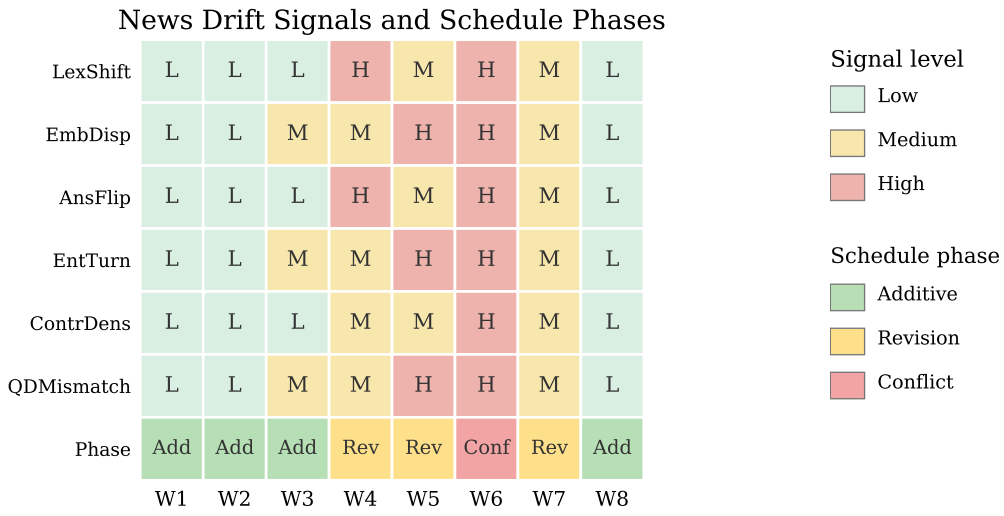

Figure 1: A News drift card showing the six frozen transition signals and their mapping to additive, revision, and conflict schedule phases. Phase labels determine ordering; the separate high/medium/null response slices are defined by s_t.

flip divides aligned training queries whose gold passage changes by all aligned training queries at the transition. Entity turnover is the symmetric difference of persistent keys divided by their union. Contradiction density divides aligned keys with incompatible current and stale values by all keys observed in both windows. Query–document mismatch is the change in cosine distance between each aligned training query and its gold passage under the same frozen encoder, averaged over aligned queries. These denominators keep corpus size from acting as a drift proxy.

Each component is winsorized at the 1st and 99th percentiles of its stream's disjoint development transitions, mapped through that development empirical CDF using midranks for ties, and clipped to $[0, 1]$. The resulting $\widetilde{d}_{t,j}$ values define

$$s_t = \frac{1}{6} \sum_{j=1}^{6} \widetilde{d}_{t,j}. \tag{4}$$

The response map predeclares high drift as $s_t \geq 0.25$, medium drift as $0.10 \leq s_t < 0.25$, and the null regime as $s_t < 0.10$. The first window has no predecessor and receives the zero vector; it contributes training pairs and null-slice evaluation queries but is excluded from every threshold fit. Normalization populations and clipping points are frozen before test scoring. Held-out evaluation queries, their relevance labels, and labels from later corpus windows never enter card construction.

3.3 StrataOrder and falsification logic

The primary phase thresholds are fixed at $\tau_a = 0.15$ for the raw answer-flip rate and $\tau_c = 0.25$ for the raw contradiction-density rate; these schedule thresholds are distinct from the normalized s_t cutoffs used for response-map slices. Assignment is exhaustive and uses conflict precedence: a window is conflict if contradiction density is at least τ_c (including dual-threshold cases), revision if it is below τ_c and answer flip is at least τ_a, and additive otherwise. StrataOrder presents additive windows chronologically, then revision windows, then conflict windows. Additive-phase replay cycles over additive pairs, revision-phase replay draws from additive pairs, and conflict-phase replay draws from both earlier phase pools. The occurrence identities are fixed before applying any comparator permutation.

A separate dev-τ calibration uses 12,288 aligned transition pairs sampled equally across the three disjoint development streams. It selects (τ_a, τ_c) to maximize stream-balanced development Conflict Preference subject to no more than one point of development forgetting and chooses 0.17 and

Table 2: Temporal data and diagnostic sizes. C/F/L gives conflict pairs, Window-1 forgetting queries, and future-entity leakage tests.

Stream	Win.	Passages	Train pairs	Dev queries	Test queries	C/F/L
News	8	2.4M	184,320	12,288	3,996	2638/1214/2312
API Docs	6	380K	96,768	8,064	3,508	2148/1036/1986
Wiki Entities	8	1.8M	164,864	10,752	4,306	2476/1188/2461

0.23. These values are frozen only for the dev-τ variant; primary STRATAORDER retains the fixed $(0.15, 0.25)$ thresholds. Reverse tests the direction of the complete ordering rule; Shuffled-Within-Phase tests the value of chronology after phase membership is preserved; dev-τ tests calibration stability; low-drift windows locate the response floor; and Time-Conditioned measures a parameterized temporal alternative under random presentation.

4 EXPERIMENTAL PROTOCOL

4.1 STREAMS AND TEMPORAL SPLITS

We evaluate three English streams whose source metadata fixes temporal membership before model training: publication timestamps form eight quarterly Reuters News windows, release tags form six versions of API documentation pages, and snapshot timestamps form eight monthly Wiki Entities states linked by Wikidata identifiers. The unit carried across windows is a persistent event–relation, API symbol–behavior, or entity–property key. Each key is assigned once to train, development, or test before pair generation, preventing versions of the same fact or behavior from crossing splits. Exact duplicate passages are collapsed within a release; changed text retains its cross-release identity. The streams contain 2.4M, 380K, and 1.8M passages, respectively; Table 2 reports every split denominator.

Adjacent-window examples are aligned only when the persistent key is unchanged. A retained evaluation item must map that key to one unambiguous current value or behavior and one release-state gold passage; missing, multi-valued, or otherwise ambiguous mappings are excluded. A gold change requires the current passage to express a value or behavior absent from, or incompatible with, its predecessor. Contradiction pairs keep the key fixed and pair the current gold with the strongest superseded passage. Card statistics and hard negatives use training keys after split assignment. Held-out query text and relevance labels never participate in threshold selection, drift-card fitting, or mining.

Future-Window Recall@100 evaluates Q_t against a release-correct historical index I_t containing every passage published no later than W_t. Superseded passages remain searchable and are explicitly nonrelevant; passages first released after t are absent. The sole relevant target is the release-state gold, so recall measures whether the current state survives competition from available stale versions. "Future-window" refers to the gold state relative to those confounders, not to hidden training data. Conflict Preference isolates the same distinction in a current-versus-superseded two-candidate comparison. Forgetting uses a method-independent reference: within each seed and stream, a shared checkpoint trained only on W_1 is evaluated on the fixed W_1 queries and compared with every final checkpoint. Leakage instead removes release masking by searching the full-history index; for a query anchored before an entity or symbol first appears, any later-first-appearance passage in the top 100 is a leakage event. This deliberate candidate-set change explains the nonzero counts in Tables 4 and 9.

4.2 MODELS, BASELINES, AND OPTIMIZATION

The primary retriever is Contriever-MSMARCO (Izacard et al., 2022); architecture transfer uses ColBERTv2 (Santhanam et al., 2022). Both retain their published initialization and native scoring function and use the independent-occurrence objective defined above. Every comparison uses the same stream splits, 50K AdamW steps, batch size 128, and learning rate 2×10^{-5} over five matched seeds. Runs use eight NVIDIA A100 80GB GPUs; the software stack is CUDA 12.4 with PyTorch

2.3.1. Within a seed, order comparisons share initialization, occurrence IDs, negative assignments, and optimizer hyperparameters.

Time-Conditioned prepends separate learned query and passage window tokens to the same seeded global permutation used by Random. Eight query and eight passage embeddings of width 768 add 12,288 parameters; the table is sized to the maximum eight-window stream, and API Docs uses six IDs while the remaining two are inactive. Vectors are initialized from $\mathcal{N}(0, 0.02)$. The target release ID is available during training and inference, and cross-window queries receive the ID of the corpus being searched. It shares Random's 20% replay occurrence set, negative assignments, five-seed budget, and selected learning rate.

Replay-only follows chronological windows with reservoir replay fixed at 20%. EWC-only follows chronological windows without replay and applies a diagonal Fisher penalty estimated from 10,000 training pairs. Development tuning selects $\lambda = 10$ from $\{0.1, 1, 10, 100\}$. Replay+EWC combines the fixed 20% reservoir with the same Fisher estimator and selects $\lambda = 1$ on the identical grid. Latest-Only trains on the current window allocation at each phase. These practical baselines receive the same total update and hyperparameter-selection budget.

4.3 Uncertainty and aggregation

Reported $\pm$ values are 95% t intervals for the mean across five matched seeds. Pairwise differences use a nested bootstrap with 10,000 replicates: seeds are resampled as matched units, evaluation-query pairs are resampled within each stream and window, and stream-level means are combined with equal weight. The resulting intervals quantify training-seed and held-out-query uncertainty conditional on the 22-window evaluation panel; windows define the response strata and aggregation units rather than an assumed random sample of all possible temporal transitions. For regime G, recall is window-balanced within each stream and stream-balanced across domains:

$$\bar{Y}_G = \frac{1}{3}\sum_{s=1}^{3} \frac{1}{|W_{s,G}|} \sum_{t \in W_{s,G}} Y_{s,t}.$$

The all-window scores in Table 6 instead micro-average queries within each stream, $Y_s^{\text{micro}} = \sum_t n_{s,t} Y_{s,t} / \sum_t n_{s,t}$, before taking the three-stream mean. Hence Table 6 is not a convex combination of the window-balanced regime means. Cohen's d divides the mean seed-paired stream-level difference by the standard deviation of those differences. Definitions, weights, and confidence procedures are fixed before comparison. The null slice contains six low-drift transitions and the three first-window zero-context controls defined in Section 3; high and medium consist only of adjacent transitions.

The evaluation first measures $\mathcal{R}$, then attributes its nonzero region with destructive sequence controls, and finally tests transport across practical comparators, severity definitions, retrieval architectures, and update budgets.

5 Results

5.1 The drift–order response map

Order response rises with transition severity on the evaluated panel. StrataOrder gains $+4.6$ over Random in high drift ($[3.1, 6.2]$, $d = 0.89$), and its paired advantage over Time-Conditioned is +2.3 ($[0.7, 3.8]$). All three streams contribute positive high-drift point estimates (Appendix A). The response contracts to +1.8 [0.7,3.0] over Random and +0.9 [-0.1,1.9] over Time-Conditioned at medium drift, then reaches $+0.3$ in the null slice. Against tuned Replay+EWC, StrataOrder retains a high-drift advantage of +3.0 [1.5,4.5]. Thus the drift card identifies the observed windows in which fixed-exposure order changes retrieval most.

5.2 Attributing the response to sequence

Holding every occurrence and per-example objective fixed, the forward sequence adds $+3.1$ R@100, $+7.0$ Conflict Preference, and 4.6 points of retained Window-1 recall relative to Random. Reverse

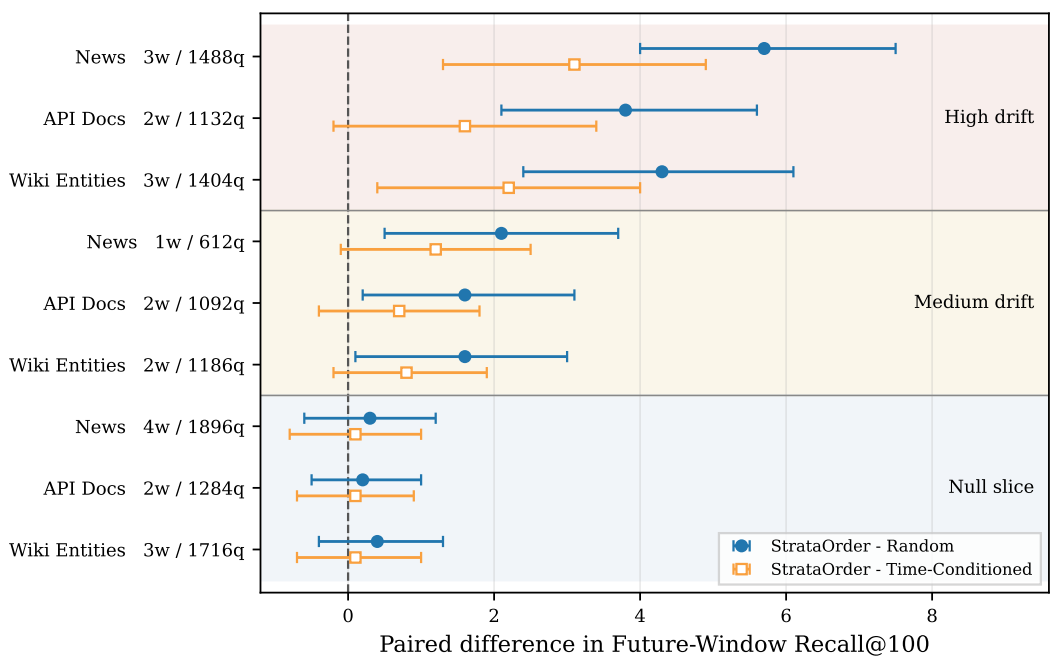

Figure 2: The drift–order response map on the fixed 22-window panel. Points are paired Recall@100 effects of StrataOrder relative to Random and Time-Conditioned; bars are 95% intervals, with window/query counts labeled. Every stream contributes to the high-drift response, medium effects contract, and null-slice effects approach zero.

Table 3: Stream-balanced Recall@100 by drift regime. Differences and 95% intervals are paired against Random; d uses seed-paired stream effects. High, medium, and null slices contain 8/4,024, 5/2,890, and 9/4,896 windows/queries.

Regime	Method	R@100	Δ_{Rnd} / 95% CI	d
High	Random	70.1 ± 1.1	—	—
	Replay-only	71.4 ± 0.9	+1.3 / [0.1, 2.6]	—
	EWC-only	70.9 ± 1.0	+0.8 / [-0.5, 2.1]	—
	Replay+EWC	71.7 ± 0.9	$+1.6/[0.3, 2.9]$	0.31
	Time-Conditioned	72.4 ± 0.8	$+2.3/[0.9, 3.7]$	0.44
	StrataOrder	74.7 ± 0.7	$+4.6/[3.1, 6.2]$	0.89
Medium	Random	73.9	—	—
	Time-Conditioned	74.8	—	—
	StrataOrder	75.7	+1.8 [0.7, 3.0]	—
Null	Random	76.1 ± 0.7	—	—
	Replay-only	76.4 ± 0.6	+0.3 / [-0.4, 1.0]	—
	EWC-only	76.2 ± 0.7	+0.1 / [-0.6, 0.8]	—
	Replay+EWC	76.5 ± 0.6	$+0.4/[-0.3, 1.1]$	0.10
	Time-Conditioned	76.3 ± 0.7	$+0.2/[-0.5, 0.9]$	0.05
	StrataOrder	76.4 ± 0.6	$+0.3/[-0.4, 1.1]$	0.08

moves all three outcomes adversely. Phase-preserving shuffling is $+1.4$ R@100 above Random, and restoring chronology inside phases adds a further descriptive $+1.7$. This ordered response establishes that the permutations are not exchangeable and links the recall gain to better current-over-stale ranking and retention. Disjointly calibrated thresholds remain within 0.2 points of the primary result.

Table 4: Matched controls on News. Every row uses the identical triplet-occurrence multiset, replay quota, initialization, negatives, and optimizer budget. Diagnostics contain 2638 conflict pairs, 1214 forgetting queries, and 2312 leakage tests.

Control	Future R@100	Conflict Pref.	Forgetting	Leakage
Random	74.2 ± 0.8	62.4 ± 1.1	8.7 ± 0.9	0.04% (1/2312)
Reverse	73.1 ± 0.9	59.8 ± 1.3	11.2 ± 1.1	0.04% (1/2312)
Shuffled-Within-Phase	75.6 ± 0.7	64.9 ± 1.0	6.8 ± 0.7	0.00% (0/2312)
STRATAORDER	77.3 ± 0.5	69.4 ± 0.8	4.1 ± 0.5	0.00% (0/2312)
STRATAORDER (dev-τ)	77.1 ± 0.6	68.9 ± 0.9	4.4 ± 0.6	0.00% (0/2312)

Table 5: ColBERTv2 replication with the same windows, triplet occurrences, replay, 50K updates, queries, and five matched seeds. Panel A repeats the News controls; Panel B reports stream-balanced regimes.

Panel A: News matched-order controls

Control	Future R@100	Conflict Pref.	Forgetting	Leakage
Random	76.8 ± 0.7	64.1 ± 1.0	8.0 ± 0.8	0.04% (1/2312)
Reverse	75.7 ± 0.8	61.4 ± 1.2	10.5 ± 1.0	0.04% (1/2312)
Shuffled-Within-Phase	77.9 ± 0.6	67.2 ± 0.9	6.3 ± 0.6	0.00% (0/2312)
STRATAORDER	79.5 ± 0.5	71.1 ± 0.8	3.8 ± 0.5	0.00% (0/2312)
STRATAORDER (dev-τ)	79.2 ± 0.6	70.6 ± 0.8	4.0 ± 0.5	0.00% (0/2312)

Panel B: three-stream regime transfer

Slice	Random	Time-Cond.	STRATAORDER	SO–Rnd 95% CI	SO–TC 95% CI
High	72.8 ± 1.0	74.6 ± 0.8	76.8 ± 0.7	+4.0 [2.5, 5.5]	+2.2 [0.6, 3.7]
Null	78.5 ± 0.7	78.7 ± 0.6	78.8 ± 0.6	+0.3 [-0.4, 1.0]	—

At the reported denominator, the forgetting reduction retains approximately 56 additional Window-1 queries; the high-drift effect corresponds to 46 additional successful retrievals per 1,000 queries.

5.3 TRANSFER ACROSS RETRIEVAL ARCHITECTURES

Table 5 shows that ColBERTv2 shifts absolute retrieval quality upward while preserving the complete directional ordering on News and the high/null boundary across streams. Its high-drift effects are +4.0 [2.5, 5.5] over Random and +2.2 [0.6, 3.7] over Time-Conditioned, with 3/3 positive stream effects across the same 8 windows and 4,024 queries. The null effect remains +0.3 [-0.4, 1.0]. The response therefore transports from single-vector Contriever geometry to ColBERTv2's late interaction.

5.4 AVERAGE PERFORMANCE AND COMPARATOR CLOSURE

STRATAORDER ranks first on every stream and improves the stream average by 1.5 points over Time-Conditioned. Replay-only, EWC-only, and their tuned hybrid cover direct replay, regularization, and combined adaptation; each remains below the structured sequence in high drift. Reverse falling below Random in all three domains supplies a second stream-wide signature of directionality. The higher News score here than in any displayed News regime mean follows from query-micro weighting inside the stream, whereas Tables 3 and 8 weight windows equally before combining streams.

5.5 SEVERITY AND THRESHOLD ROBUSTNESS

Stricter high-drift selection increases effect magnitude as the slice contracts, from +4.0 [2.7,5.3] at $s_t \geq 0.20$ to +5.0 [3.3,6.8] at $s_t \geq 0.30$. Both neighboring null definitions remain centered near zero. The alternate card averages its three normalized signals and applies the primary $0.25/0.10$ cutoffs; it retains +3.8 [2.2,5.4] in high drift and +0.2 [-0.5,0.9] in the null slice. The response therefore tracks observable change across label composition and threshold choices.

Table 6: Query-micro Future-Window Recall@100 within each stream under matched 50K-step budgets; Average weights streams equally. Intervals are 95% across five seeds, and stars mark column maxima.

Method	News	API Docs	Wiki Ent.	Average
Random	74.2 ± 0.8	71.8 ± 0.9	73.1 ± 0.7	73.0 ± 0.6
Latest-Only	74.8 ± 0.7	72.4 ± 0.8	73.6 ± 0.7	73.6 ± 0.5
Replay-only	75.2 ± 0.7	73.1 ± 0.8	74.1 ± 0.6	74.1 ± 0.5
EWC-only	74.9 ± 0.7	72.9 ± 0.8	73.8 ± 0.7	73.9 ± 0.6
Replay+EWC	75.4 ± 0.6	73.6 ± 0.7	74.2 ± 0.6	74.4 ± 0.5
Time-Conditioned	76.1 ± 0.6	74.2 ± 0.7	74.8 ± 0.6	75.0 ± 0.5
Reverse	73.1 ± 0.9	70.9 ± 1.0	72.4 ± 0.8	72.1 ± 0.7
StrataOrder	$77.3^{\star} \pm 0.5$	$75.8^{\star} \pm 0.6$	$76.4^{\star} \pm 0.5$	$76.5^{\star} \pm 0.4$

Table 7: Sensitivity of paired Recall@100 effects. Counts are windows/queries; the alternate card uses only lexical shift, embedding displacement, and entity turnover normalized on disjoint development data.

Card	Slice	Count	SO–Random 95% CI	SO–Time-Cond. 95% CI
Primary	$s_t \geq 0.20$	11 / 5512	+4.0 [2.7, 5.3]	+1.9 [0.5, 3.3]
Primary	$s_t \geq 0.25$	8 / 4024	$+4.6[3.1, 6.2]$	$+2.3[0.7, 3.8]$
Primary	$s_t \geq 0.30$	6 / 3010	+5.0 [3.3, 6.8]	+2.6 [0.8, 4.3]
Primary	$s_t < 0.08$	7 / 3820	+0.2 [-0.5, 0.9]	—
Primary	$s_t < 0.10$	9 / 4896	$+0.3[-0.4, 1.1]$	—
Primary	$s_t < 0.12$	11 / 6038	+0.4 [-0.3, 1.1]	—
Alternate	high	7 / 3568	+3.8 [2.2, 5.4]	+1.8 [0.3, 3.4]
Alternate	null	8 / 4372	+0.2 [-0.5, 0.9]	—

6 Discussion and Implications

The response map changes how temporal retriever refresh should treat scheduling. Under a fixed presentation multiset and independent-occurrence loss, revision/conflict windows show a large response while the null slice remains near zero; time inputs, replay/EWC variants, two retrieval geometries, threshold changes, and an alternate card preserve this boundary. The ladder Reverse $<$ Random $<$ Shuffled-Within-Phase $<$ StrataOrder links recall to current-over-stale preference and retention, and the 25K–75K sweep rules out a single-horizon effect. A stream-wide average would hide the windows in which sequence design matters most.

The identified object is the complete StrataOrder rule for retrospective full-history refresh with 20% replay across three English streams and two dense-retriever architectures. The tested schedules jointly encode phase progression, chronology, and terminal position, so the mechanistic interpretation attaches to the complete rule. The null control comprises six low-drift transitions plus three zero-context first windows. These boundaries make rolling refresh, additional languages, and generative retrievers precise tests of the same matched-sequence estimand.

7 Conclusion

Training sequence changes temporal retrieval even when 6.4 million triplet presentations, replay identities, per-example losses, initialization, optimizer configuration, and the learning-rate-by-step schedule are matched. Forward additive–revision–conflict order yields a $+4.6$-point high-drift gain, contracts at medium drift, and approaches the response floor in the null slice; ColBERTv2 and three optimization budgets preserve the boundary. The matched-multiset response map therefore provides both an identification result and a decision signal: retrospective refresh benefits most from deliberate scheduling where revisions and contradictions change the state encoded by otherwise relevant passages.

AI USE STATEMENT

Generative AI tools were used only for manuscript writing, editing, and presentation. The authors reviewed all edits and remain responsible for the final content.

REFERENCES

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11849–11860. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/15825aee15eb335cc13f9b559f166ee8-Abstract.html.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/b704ea2c39778f07c617f6b7ce480e9e-Abstract.html.

Yinqiong Cai, Keping Bi, Yixing Fan, Jiafeng Guo, Wei Chen, and Xueqi Cheng. L2R: Lifelong learning for first-stage retrieval with backward-compatible representations. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 183–192. ACM, 2023. doi: 10.1145/3583780.3614947. URL https://doi.org/10.1145/3583780.3614947.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hkf2_sC5FX.

Dipam Goswami, Liying Wang, Bartłomiej Twardowski, and Joost van de Weijer. Query drift compensation: Enabling compatibility in continual learning of retrieval embedding models. In *Proceedings of the 4th Conference on Lifelong Learning Agents*, volume 330 of *Proceedings of Machine Learning Research*, pp. 324–340. PMLR, 2026. URL https://proceedings.mlr.press/v330/goswami26a.html.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2535–2544. PMLR, 2019. URL https://proceedings.mlr.press/v97/hacohen19a.html.

Xingwei He, Yeyun Gong, A-Long Jin, Hang Zhang, Anlei Dong, Jian Jiao, Siu Yiu, and Nan Duan. CAPSTONE: Curriculum sampling for dense retrieval with document expansion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10531–10541. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.651. URL https://aclanthology.org/2023.emnlp-main.651/.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=jKN1pXi7b0.

Joel Jang, Seonghyeon Ye, Changho Lee, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, and Minjoon Seo. TemporalWiki: A lifelong benchmark for training and evaluating ever-evolving language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6237–6250. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.418. URL https://aclanthology.org/2022.emnlp-main.418/.

Yaning Jia, Chunhui Zhang, Xingjian Diao, Xiangchi Yuan, Zhongyu Ouyang, Chiyu Ma, and Soroush Vosoughi. What makes a good curriculum? disentangling the effects of data ordering on LLM mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34472–34488. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1591. URL https://aclanthology.org/2026.acl-long.1591/.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2304–2313. PMLR, 2018. URL https://proceedings.mlr.press/v80/jiang18c.html.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Jungo Kasai, Keisuke Sakaguchi, Yoichi Takahashi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. RealTime QA: What's the answer right now? In *Advances in Neural Information Processing Systems*, volume 36, pp. 49025–49043. Curran Associates, Inc., 2023. doi: 10.52202/075280-2130. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9941624ef7f867a502732b5154d30cb7-Abstract-Datasets_and_Benchmarks.html.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL https://doi.org/10.1145/3397271.3401075.

Chaeeun Kim, Soyoung Yoon, Hyunji Lee, Joel Jang, Sohee Yang, and Minjoon Seo. Exploring the practicality of generative retrieval on dynamic corpora. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13616–13633. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.755. URL https://aclanthology.org/2024.emnlp-main.755/.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/10.1073/pnas.1611835114.

Dayoon Ko, Jinyoung Kim, Sohyeon Kim, Jinhyuk Kim, Jaehoon Lee, Seonghak Song, Minyoung Lee, and Gunhee Kim. When should dense retrievers be updated in evolving corpora? detecting out-of-distribution corpora using GradNormIR. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25977–25996. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1334. URL https://aclanthology.org/2025.findings-acl.1334/.

M. Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, volume 23, pp. 1189–1197. Curran Associates, Inc., 2010. URL https://proceedings.neurips.cc/paper/2010/hash/e57c6b956a6521b28495f2886ca0977a-Abstract.html.

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 for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Adam Liska, Tomas Kocisky, Elena Gribovskaya, Tayfun Terzi, Eren Sezener, Devang Agrawal, Cyprien De Masson D'Autume, Tim Scholtes, Manzil Zaheer, Susannah Young, Ellen Gilsenan-Mcmahon, Sophia Austin, Phil Blunsom, and Angeliki Lazaridou. StreamingQA: A benchmark for adaptation to new knowledge over time in question answering models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 13604–13622. PMLR, 2022. URL https://proceedings.mlr.press/v162/liska22a.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6467–6476. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/f87522788a2be2d171666752f97ddebb-Abstract.html.

Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. DSI++: Updating transformer memory with new documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8198–8213. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.510. URL https://aclanthology.org/2023.emnlp-main.510/.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1162–1172. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1119. URL https://aclanthology.org/N19-1119/.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5835–5847. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.466. URL https://aclanthology.org/2021.naacl-main.466/.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5533–5542. IEEE, 2017. doi: 10.1109/CVPR.2017.587. URL https://doi.org/10.1109/CVPR.2017.587.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, pp. 350–360. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.272. URL https://aclanthology.org/2022.naacl-main.272/.

HuiJeong Son, Hyeongu Kang, Sunho Kim, Subeen Ho, SeongKu Kang, Dongha Lee, and Susik Yoon. CREAM: Continual retrieval on dynamic streaming corpora with adaptive soft memory. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1297–1308. ACM, 2026. doi: 10.1145/3770854.3780281. URL https://doi.org/10.1145/3770854.3780281.

Si Sun, Chenyan Xiong, Yue Yu, Arnold Overwijk, Zhiyuan Liu, and Jie Bao. Reduce catastrophic forgetting of dense retrieval training with teleportation negatives. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6639–6654. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.445. URL https://aclanthology.org/2022.emnlp-main.445/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, pp. 359–370, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13697–13720. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.813. URL https://aclanthology.org/2024.findings-acl.813/.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=zeFrfgyZln.

Hansi Zeng, Hamed Zamani, and Vishwa Vinay. Curriculum learning for dense retrieval distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1979–1983. ACM, 2022. doi: 10.1145/3477495.3531791. URL https://doi.org/10.1145/3477495.3531791.

A Per-Stream Regime Decomposition

Table 8 gives the complete primary-backbone decomposition behind Figure 2. Every stream contributes a positive high-drift contrast against both Random and Time-Conditioned. Medium-drift point estimates remain positive and smaller, with the stream-balanced Random contrast clearing zero. The null panel contains three first-window zero-context controls and six low-drift adjacent transitions; its estimates cluster tightly around zero.

Table 8: Per-stream regime map for Contriever-MSMARCO. Entries are Recall@100; differences include paired 95% intervals. Counts are windows/evaluation queries.

Stream	Slice	Random	Time-Cond.	StrataOrder	SO–Rnd	SO–TC
News	High	69.4	72.0	75.1	+5.7 [4.0, 7.5]	+3.1 [1.3, 4.9]
API Docs	High	70.6	72.8	74.4	+3.8 [2.1, 5.6]	+1.6 [-0.2, 3.4]
Wiki Ent.	High	70.3	72.4	74.6	+4.3 [2.4, 6.1]	+2.2 [0.4, 4.0]
News	Medium	73.5	74.4	75.6	+2.1 [0.5, 3.7]	+1.2 [-0.1, 2.5]
API Docs	Medium	73.9	74.8	75.5	+1.6 [0.2, 3.1]	+0.7 [-0.4, 1.8]
Wiki Ent.	Medium	74.3	75.1	75.9	+1.6 [0.1, 3.0]	+0.8 [-0.2, 1.9]
News	Null	75.8	76.0	76.1	+0.3 [-0.6, 1.2]	+0.1 [-0.8, 1.0]
API Docs	Null	76.4	76.5	76.6	+0.2 [-0.5, 1.0]	+0.1 [-0.7, 0.9]
Wiki Ent.	Null	76.1	76.4	76.5	+0.4 [-0.4, 1.3]	+0.1 [-0.7, 1.0]

High-slice counts are 3 / 1488, 2 / 1132, and 3 / 1404 for News, API Docs, and Wiki Entities. Medium counts are 1 / 612, 2 / 1092, and 2 / 1186; null counts are 4 / 1896, 2 / 1284, and 3 / 1716. These sum to the stream test sizes and preserve the stream-balanced aggregate.

B Matched-Control Replication Beyond News

The complete order-control ladder is repeated on both non-News streams with the same diagnostic definitions and matched-seed protocol.

Table 9: Matched-order controls on API Docs and Wiki Entities; SO and SWP abbreviate StrataOrder and Shuffled-Within-Phase. API diagnostics contain 2148/1036/1986 conflict/forgetting/leakage cases; Wiki diagnostics contain 2476/1188/2461.

Stream	Control	Future R@100	Conflict Pref.	Forgetting	Leakage
API	Random	71.8 ± 0.9	61.7 ± 1.2	9.4 ± 0.8	0.05% (1/1986)
API	Reverse	70.9 ± 1.0	59.1 ± 1.3	11.6 ± 1.0	0.05% (1/1986)
API	SWP	74.1 ± 0.8	64.0 ± 1.0	7.1 ± 0.7	0.00% (0/1986)
API	StrataOrder	75.8 ± 0.6	67.1 ± 0.8	4.9 ± 0.5	0.00% (0/1986)
API	SO (dev-τ)	75.5 ± 0.7	66.6 ± 0.9	5.1 ± 0.5	0.00% (0/1986)
Wiki	Random	73.1 ± 0.7	62.9 ± 1.0	8.1 ± 0.7	0.04% (1/2461)
Wiki	Reverse	72.4 ± 0.8	60.4 ± 1.1	10.3 ± 0.8	0.04% (1/2461)
Wiki	SWP	74.7 ± 0.7	65.3 ± 0.9	6.2 ± 0.6	0.00% (0/2461)
Wiki	StrataOrder	76.4 ± 0.5	68.6 ± 0.7	4.3 ± 0.4	0.00% (0/2461)
Wiki	SO (dev-τ)	76.2 ± 0.6	68.1 ± 0.8	4.6 ± 0.5	0.00% (0/2461)

Both non-News streams reproduce the full ordering: Reverse falls below Random, phase-preserving shuffling is intermediate, StrataOrder is highest, and disjoint-development calibration remains close to the primary thresholds. Conflict Preference and forgetting move coherently with future recall.

C COLBERTV2 PER-STREAM TRANSFER

Table 10 repeats the response-map endpoints separately for each stream, preserving the original windows and evaluation queries.

Table 10: ColBERTv2 per-stream regime transfer. Differences include paired 95% intervals and use the same windows/queries as Table 8.

Stream	Slice	Random	Time-Cond.	STRATAORDER	SO–Rnd	SO–TC
News	High	72.2	74.5	76.7	+4.5 [2.8, 6.1]	+2.2 [0.5, 3.9]
API Docs	High	73.1	74.6	76.4	+3.3 [1.5, 5.0]	+1.8 [-0.1, 3.6]
Wiki Ent.	High	73.1	74.7	77.3	+4.2 [2.4, 5.9]	+2.6 [0.7, 4.4]
News	Null	78.2	78.4	78.5	+0.3 [-0.5, 1.1]	+0.1 [-0.6, 0.8]
API Docs	Null	78.8	78.9	79.0	+0.2 [-0.5, 0.9]	+0.1 [-0.6, 0.8]
Wiki Ent.	Null	78.5	78.8	78.9	+0.4 [-0.3, 1.1]	+0.1 [-0.6, 0.8]

Absolute ColBERTv2 scores differ from Contriever-MSMARCO, but all three high-drift point estimates share the same sign. The stream-balanced effects clear zero against both Random and Time-Conditioned, while the null-slice estimates remain tightly centered near zero.

D BUDGET SENSITIVITY

Table 11 brackets the primary 50K comparison with shorter and longer training schedules, testing whether the high/null boundary depends on one optimization horizon.

Table 11: Paired STRATAORDER–Random effects across update budgets. News is the complete matched-control contrast; high and null are stream-balanced regime contrasts.

Updates	News 95% CI	High-drift 95% CI	Null 95% CI
25K	+2.5 [1.0, 4.0]	+3.9 [2.3, 5.5]	+0.2 [-0.6, 1.0]
50K	+3.1 [1.8, 4.4]	$+4.6[3.1, 6.2]$	$+0.3[-0.4, 1.1]$
75K	+2.8 [1.4, 4.2]	+4.2 [2.6, 5.8]	+0.4 [-0.4, 1.2]

All three budgets preserve a resolved high-drift contrast and a null-slice estimate near zero. The effect is therefore present before and after the primary stopping point rather than emerging only at 50K updates.

E TEMPORAL INTEGRITY CHECKS

Primary recall searches all passages released through boundary t, retaining superseded versions but excluding later releases. The leakage diagnostic intentionally exposes the full-history index and asks whether a later-first-appearance passage enters the top 100 for an earlier-anchored query. This separation tests premature ranking without contaminating the primary candidate set. Release timestamps and version tags are resolved before key-level splitting; duplicate text inside a release is collapsed, and adjacent states must share the same event–relation, symbol–behavior, or entity–property key before they can form a current/stale pair.

Drift features use training transitions only, with disjoint development transitions supplying normalization and the alternate-card check in Table 7. Replay integrity is established at the occurrence level: the matched-order family shares all 6.4 million occurrence IDs, replay designations, multiplicities, and hard negatives before any permutation. Tables 4 and 9 then evaluate the same first-appearance, conflict, and forgetting query sets across methods.