SmallReason-ColBERT: An Ultra-Small Late-Interaction Retriever for Reasoning Intensive Retrieval

Abdelrahman Abdallah, Mohammed Ali, Adam Jatowt
University of Innsbruck, Austria
{abdelrahman.abdallah,adam.jatowt}@uibk.ac.at

Abstract

Reasoning-intensive retrieval remains difficult for small models. Compact public ColBERTs are usually trained on general-purpose corpora and underperform reasoning-tuned 150M+ baselines on BRIGHT (Su et al., 2025) by several nDCG@10 points. However, no public reasoning-tuned ColBERT exists at edge scale. We introduce **SmallReason-ColBERT**, a 32M late-interaction retriever that closes much of this gap with three components: a varied-length contrastive warmup on ReasonIR-VL, a hard-negative contrastive polish on merged ReasonIR-HQ and BGE-Reasoner data, and a single-layer per-query-token importance head trained on top of the frozen base. The head is trained with an un-normalised weighted MaxSim score and evaluated with its length-normalised form. In a controlled re-training, replacing this training objective with the symmetric normalised score causes the loss to stall and costs 3.59 nDCG@10. The full recipe reaches **21.41** mean nDCG@10 on BRIGHT, within 1.21 of the 150M Reason-ModernColBERT (22.62) and above all $\leq$ 33M ColBERTs we evaluate. Through ablations on capacity, initialisation, and score variants, we further show that the learned head outperforms fixed IDF weighting and that simply thresholding the learned gates is harmful.[1]

1 Introduction

Retrieval is now a central component of language-model systems, search engines, and domain-specific question answering pipelines (Croft et al., 2010; Abdallah et al., 2026c; Herzig et al., 2021). As these systems are asked to support longer, more technical, and more reasoning-heavy queries, the retrieval problem increasingly goes beyond lexical overlap or broad semantic similarity: a useful retriever must identify documents that match the intent, constraints, and intermediate concepts of a query (Abdallah et al., 2026a). This shift has exposed a gap between efficient retrievers that can be deployed widely and larger retrieval models tuned for reasoning-intensive benchmarks.

Late-interaction retrievers (Khattab and Zaharia, 2020; Santhanam et al., 2022) have become a competitive alternative to single-vector dense retrieval. By scoring queries and documents at the granularity of individual tokens and combining them with a per-token MaxSim aggregator, they capture fine-grained interactions that bi-encoders miss while keeping inference costs close to those of bi-encoders. Two recent directions have pushed the family further: small encoders that target on-device retrieval (Takehi et al., 2025; Clavié, 2024), and reasoning-tuned encoders that target the new generation of reasoning-intensive benchmarks (Chaffin, 2025b; Su et al., 2025). These directions have not been combined. Public reasoning-tuned ColBERT models are still built at the 150M-parameter scale (Chaffin, 2025b), while compact public ColBERTs are trained as general-purpose retrievers (Takehi et al., 2025; Clavié, 2024). Practitioners who need reasoning-aligned retrieval at edge scale therefore face an awkward choice: use a larger reasoning-tuned retriever, or use a compact retriever whose training objective was not designed for reasoning-heavy queries.

In this paper we attempt to close that gap with a three-step recipe applied to a public 32M ColBERT. We first run a varied-length contrastive *warmup* on ReasonIR-VL (Shao et al., 2025) to align the model with reasoning-style queries of mixed length. We then *polish* the resulting model with a hard-negative contrastive stage on a merged ReasonIR-HQ and BGE-Reasoner (Chen et al., 2025) corpus. The first two steps produce a reasoning-tuned base, but on their own they do not match larger reasoning ColBERTs on BRIGHT. The third and final step adds a single-layer per-query-token *importance head* on

[1] https://github.com/DataScienceUIBK/SmallReason-ColBERT

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 10564–10575
October 24-29, 2026 ©2026 Association for Computational Linguistics

top of the frozen base. The head is small but its training is non-trivial: it must be trained against an un-normalised weighted MaxSim and evaluated with the length-normalised score. Replacing this asymmetric training-evaluation pair with the symmetric normalised score collapses the cross-entropy gradient, the head fails to depart from its initialisation, and the controlled BRIGHT mean drops by 3.59 nDCG@10. The asymmetry is the single most consequential design lever in the recipe.

Contributions. We make the following contributions:

- We introduce **SmallReason-ColBERT**, a 32M late-interaction retriever produced by a three-step recipe: warmup, hard-negative polish, and an importance head. The full recipe reaches **21.41** mean nDCG@10 on BRIGHT, within 1.21 of the 150M ReasonModernColBERT and substantially above all $\leq$ 33M ColBERTs we evaluate.
- We identify a training objective for the head that uses an un-normalised weighted MaxSim score while retaining length-normalised scores at evaluation time. With this objective the head adds $+1.80$ nDCG@10 over the reasoning-tuned base; with the normalised training score, the head fails to learn useful gates. The recipe is robust to its training hyperparameters across an 18-cell sweep (range 0.42 nDCG@10).
- We provide a focused study of score variants on a frozen base. Adding capacity to the head hurts, fixed IDF weighting underperforms the learned gate, and thresholding the learned gates discards useful soft weighting. The single linear head is the best query-weighting variant we tested at this scale.
- We characterise what the head learns. Contrary to the "soft-IDF" interpretation of token importance gates, the learned gate values have near-zero rank correlation with token IDF (Spearman $\rho = -0.02$); the head amplifies tokens that mark query intent rather than tokens that are rare in the corpus.

2 Related Work

Late interaction. ColBERT (Khattab and Zaharia, 2020) introduced the per-token MaxSim aggregator we build on; ColBERTv2 (Santhanam et al., 2022) added denoised supervision and residual compression that have since become standard. JaColBERTv2.5 (Clavié, 2025) explored normalisation choices in the distillation loss, including options closely related to but distinct from the train-eval asymmetry we use. Clavié et al. (2025) studied the late-interaction projection head and motivated our investigation of single-vs-multi-layer head choices in §6.4.

Compact ColBERT models. The most directly comparable systems are `answerai-colbert-small-v1` (Clavié, 2024), a 33M BERT-MiniLM-based ColBERT, and the `mxbai-edge-colbert-v0` family (Takehi et al., 2025) at 17M and 32M parameters built on Ettin (Weller et al., 2025). Both are trained on classical IR data and target on-device deployment; neither is reasoning-tuned. Our base re-uses the mxbai-edge backbone but replaces its training distribution and adds the importance head on top.

ModernBERT-based late interaction. GTE-ModernColBERT (Chaffin, 2025a) is the standard 150M ColBERT built on the ModernBERT family (Warner et al., 2025); ReasonModernColBERT (Chaffin, 2025b) fine-tunes it on ReasonIR-HQ for reasoning-heavy retrieval and is our closest qualitative reference. Our 32M model approaches its BRIGHT performance with a $4.7\times$ smaller backbone and a head of only 129 parameters.

Reasoning-heavy retrieval. BRIGHT (Su et al., 2025), RECOR (Ali et al., 2026) and MM-bright (Abdallah et al., 2026b) introduced the long-form reasoning queries. ReasonIR (Shao et al., 2025) contributed the synthetic data we use for warmup and polish, along with the 8B reasoning-tuned single-vector retriever ReasonIR-8B (also a baseline in Table 1). BGE-Reasoner (Chen et al., 2025) is a complementary reasoning-aligned hard-negative dataset used in our polish stage. A line of work on sparse retrieval, including SPLADE (Formal et al., 2021b,a), COIL (Gao et al., 2021), and uniCOIL (Lin and Ma, 2021), learns token-level scalar weights through MLM-style heads.

Our head is conceptually related but operates inside a multi-vector late-interaction score rather than a sparse vocabulary projection. Our analysis (§7) shows that the learned weights are not soft-IDF, contrary to the typical motivation for SPLADE-family weights. To our knowledge no prior work on late interaction articulates the train-unnormalised / eval-normalised asymmetry we identify in §3.5.

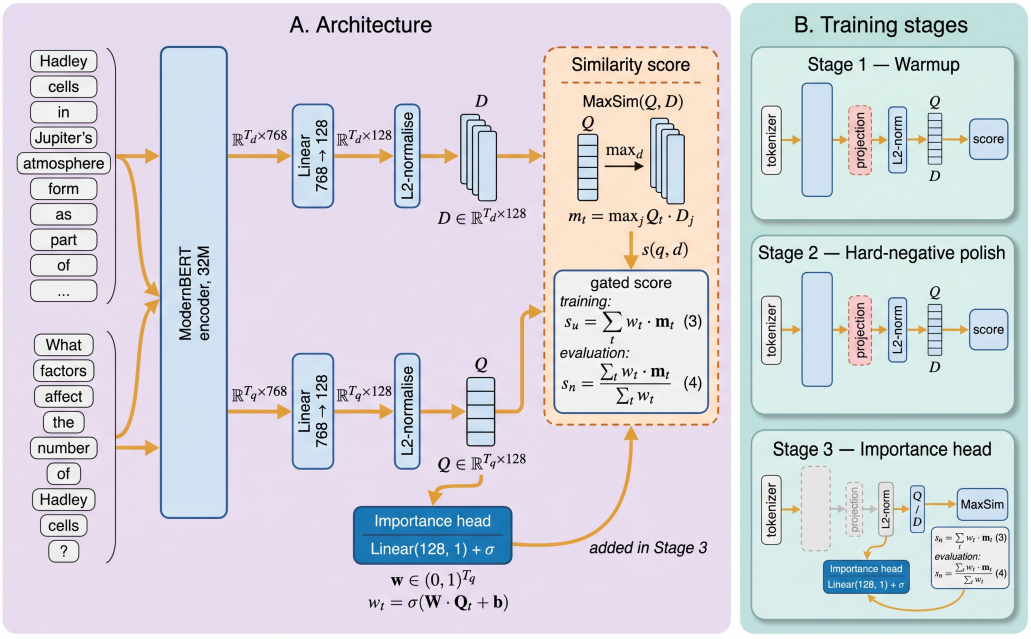

Figure 1: Overview of SmallReason-ColBERT. Left: the 32M ModernBERT-based ColBERT encodes query and document tokens, widens the late-interaction projection to 128 dimensions, and applies a query-side importance head before weighted MaxSim scoring. Right: the three-stage recipe first warms up the widened base, then polishes it with hard negatives, and finally trains only the importance head on top of the frozen base.

3 Method

Figure 1 summarises the architecture and the three training stages. The method has two parts: first we construct a reasoning-tuned 32M ColBERT base by widening and training the projection layer, and then we freeze that base and train a 129-parameter query-side importance head for weighted late interaction.

3.1 Background: Late Interaction and MaxSim

A ColBERT (Khattab and Zaharia, 2020; Santhanam et al., 2022) retriever encodes a query q into a sequence of ℓ_2-normalised vectors $\mathbf{Q} \in \mathbb{R}^{T_q \times d}$ and a document d into $\mathbf{D} \in \mathbb{R}^{T_d \times d}$. Relevance is scored by the MaxSim aggregator:

$$s_{\text{MaxSim}}(q, d) = \sum_{t=1}^{T_q} \max_{j=1}^{T_d} \mathbf{Q}_t \cdot \mathbf{D}_j. \qquad (1)$$

Each query token contributes its best match against any document token to the sum. The score is differentiable end-to-end and decomposes additively across query tokens. Standard MaxSim also treats every query token as equally informative, which is the property we revisit in §3.4.

3.2 Starting Point and Projection Widening

We start from `mxbai-edge-colbert-v0-32m` (Takehi et al., 2025), a public 32M ColBERT built on the ModernBERT family (Warner et al., 2025; Weller et al., 2025). Its 3-layer projection emits 64-dimensional token vectors, a deliberately compact choice intended for on-device deployment. On reasoning-heavy splits where many tokens look similar at the surface (variables, operators, theorem names), 64 channels are a tight bottleneck. We therefore widen the final projection `Linear` from out_features $= 64$ to out_features $= 128$. The first 64 rows of the new weight matrix are inherited from the original; the remaining 64 are initialised from $\mathcal{N}(0, \sigma^2)$ with σ set to 10% of the empirical standard deviation of the original weight matrix. The new channels are small enough that the model's MaxSim score is approximately unchanged at step zero, while still being non-zero so that they receive gradient flow during warmup. With a zero initialisation, both query and document vectors would be zero on the new channels, making the dot product and its gradient zero on those dimensions.

3.3 Reasoning-Tuned Base Training

Going directly from the widened model to hard-negative contrastive training is unstable: the random-initialised new channels inject too much noise for a stable curriculum. We therefore train the base in two stages. First, a varied-length warmup

aligns the widened projection with reasoning-style queries of mixed length. Second, a hard-negative distillation stage polishes the model on triples that are closer to the downstream retrieval distribution. The concrete corpora, losses, and training hyperparameters are given in §4.1; this section only defines the model produced by the stages. After warmup and polish, the result is a 32M reasoning-tuned ColBERT that still uses the plain MaxSim score in Eq. 1. We use this model as the frozen base for the importance head introduced next, and refer to it as *the base*. The base's retrieval behaviour is analysed in §5.1; its role in the method is to provide reasoning-aligned token embeddings while leaving the final per-token aggregation uniform.

3.4 The Importance Head

We replace the uniform per-token sum in Eq. 1 with a learned per-token weighting. A single linear layer scores each query token's d-dimensional embedding into a sigmoid-bounded gate:

$$w_t = \sigma(\mathbf{W} \cdot \mathbf{Q}_t + b), \quad \mathbf{W} \in \mathbb{R}^{1\times d},\ b \in \mathbb{R}. \quad (2)$$

For $d = 128$ this head has 129 trainable parameters, less than 0.001% of the base. The head is the only component trained in this stage; the base is frozen.

3.5 Train Un-Normalised, Evaluate Length-Normalised

The head can be incorporated into the score in two natural ways:

$$s_{\mathrm{u}}(q,d) = \sum_{t=1}^{T_q} w_t \cdot \max_{j=1}^{T_d} \mathbf{Q}_t \cdot \mathbf{D}_j, \quad (3)$$

$$s_{\mathrm{n}}(q,d) = \frac{1}{\sum_t w_t} \sum_{t=1}^{T_q} w_t \cdot \max_{j=1}^{T_d} \mathbf{Q}_t \cdot \mathbf{D}_j. \quad (4)$$

The un-normalised score s_{u} scales linearly with the query length T_q; the length-normalised score s_{n} has the same scale across queries and lies in $[-1, 1]$.

We train against s_{u} but evaluate against s_{n}. This choice matters for optimisation. If we train against s_{n}, the per-pair score difference $|\Delta s_{\mathrm{n}}|$ is bounded by the cosine range of a single token's contribution and the cross-entropy gradient over a positive-negative pair drops to $\sim 5 \times 10^{-3}$ at typical magnitudes. The loss stalls near $\ln 2$ and the gates do not learn useful weights. Replacing the un-normalised training loss with the symmetric normalised one degrades BRIGHT mean nDCG@10 by 3.59 in the controlled ablation (§6.1). Training against s_{u} lets gradients scale with query length and gives the head a usable learning signal. Within a single query, the rescaling factor $1/\sum_t w_t$ is constant across documents, so evaluating against s_{n} does not change per-query rankings; it only keeps scores comparable across query lengths.

3.6 Initialisation: a No-Op at Step Zero

We initialise $\mathbf{W} = \mathbf{0}$ and $b = 5$, so that at step zero every gate equals $\sigma(5) \approx 0.993$. Under this initialisation s_{n} reduces to plain MaxSim up to a per-query rescaling that does not change rankings, so the head implements a no-op against the frozen base before any optimisation. This gives the head a stable starting point rather than forcing it to recover from random token weights. We ablate the choice of b in §6.2. We refer to the final retriever, the reasoning-tuned base plus the trained importance head, as **SmallReason-ColBERT**.

4 Experimental Setup

4.1 Training Data and Recipe

Warmup data. The first base-training stage uses ReasonIR-VL (Shao et al., 2025), a varied-length reasoning corpus with approximately 245K triples. The mixture contains both long-context reasoning queries and shorter retrieval pairs, making it a natural curriculum for the widened projection described in §3.2.

Polish data. The second base-training stage uses a merged hard-negative corpus of ReasonIR-HQ (Shao et al., 2025) and BGE-Reasoner-data (Chen et al., 2025), totalling about 2.7M (query, positive, negative) tuples. Both sources target reasoning-aligned retrieval with cross-encoder-mined hard negatives.

Base training. The warmup stage uses a Cached-Contrastive loss over in-batch negatives; 8× H100, batch 32 per GPU, gradient accumulation 4, learning rate 1×10^{-5}, one epoch, query length 256, document length 2,048, flash-attention 2 (Dao, 2024), and bf16. Training takes about eight hours. The polish stage uses the same CachedContrastive objective on hard negatives, with the same hardware, batch size, sequence lengths, and precision, a learning rate of 5×10^{-6}, and gradient accumulation 2. Training takes about 16 hours.

Head training. The base is frozen and only the 129-parameter importance head is trained. We use AdamW with learning rate 5×10^{-4}, batch size 16, and 3,000 steps on the merged ReasonIR-HQ +

Table 1: Main BRIGHT results: nDCG@10 ($\times 100$). Best at $\leq$33M is **bold** (ties bolded jointly); best overall on the Mean column is underlined. ColBERT family models are re-evaluated under our protocol (uniform query length 256, document length 2048); other baselines are taken from the BRIGHT release and the Reason-ModernColBERT report (Chaffin, 2025b). The Reason-ModernColBERT entry marked † is the published number with per-split tuned query lengths; under our same-protocol evaluation it scores 21.97.

Model	Params	Mean	Bio	Earth	Econ	Psych	Robo	Stack	Sust	Pony	Leet	AoPS	TheoQ	TheoT
Sparse / classical IR														
BM25	–	14.53	18.90	27.20	14.90	12.50	13.60	18.40	15.00	7.90	24.40	6.20	10.40	4.90
Single-vector dense / instruction-tuned														
BGE	567M	13.71	11.70	24.60	16.60	17.50	11.70	10.80	13.30	5.70	26.70	6.00	13.00	6.90
E5-Mistral	7B	17.93	18.60	26.00	15.50	15.80	16.30	11.20	18.10	4.90	28.70	7.10	26.10	26.80
GritLM	7B	20.95	24.80	32.30	18.90	19.80	17.10	13.60	17.80	22.00	29.90	8.80	25.20	21.20
Qwen3-Embedding	7B	22.51	30.60	36.40	17.80	24.60	13.20	22.20	14.80	9.90	25.50	14.40	27.80	32.90
ReasonIR-8B	8B	<u>24.38</u>	26.20	31.40	23.30	30.00	18.00	23.90	20.50	10.50	35.00	14.70	31.90	27.20
Multi-vector / late interaction (ColBERT family)														
GTE-ModernColBERT	150M	21.72	35.30	47.86	18.97	29.70	20.49	20.18	20.76	10.58	29.99	8.91	12.87	5.07
Reason-ModernColBERT†	150M	22.62	33.25	41.02	24.93	30.73	21.12	20.62	20.31	8.51	31.07	9.17	19.51	11.24
Compact ColBERT ($\leq$33M)														
mxbai-edge-colbert (17M)	17M	18.60	29.98	41.98	17.76	20.29	16.65	16.87	19.54	12.32	27.72	10.22	8.09	1.82
answerai-colbert-small	33M	18.49	28.92	37.99	18.64	21.09	**18.95**	16.90	18.08	12.87	27.36	8.90	8.42	**3.76**
mxbai-edge-colbert (32M)	32M	18.16	28.70	42.29	17.65	21.93	18.09	16.49	18.64	12.90	16.15	9.80	12.51	2.76
SmallReason-ColBERT (no head)	32M	19.61	33.16	44.28	**20.25**	24.91	18.65	16.66	20.11	**22.77**	17.40	4.89	9.04	3.19
SmallReason-ColBERT	32M	**21.41**	**34.17**	**45.03**	19.99	**24.94**	18.14	**17.21**	**21.07**	19.33	**29.98**	**10.29**	**13.00**	3.74

BGE-Reasoner triples. Training-time scoring uses s_{u} (Eq. 3); evaluation-time scoring uses s_{n} (Eq. 4). Head training takes roughly twelve minutes on a single H100.

4.2 Evaluation

Evaluation benchmark. We evaluate on **BRIGHT** (Su et al., 2025), a reasoning-heavy retrieval benchmark with 12 splits across coding (LeetCode, Pony), math (AoPS, TheoremQA-questions, TheoremQA-theorems), and StackExchange domains (Biology, Earth Science, Economics, Psychology, Robotics, StackOverflow, Sustainable Living). All numbers are mean nDCG@10 ($\times 100$).

Eval protocol. We use a uniform query length of 256 across all splits except Pony (32) and a document length of 2048 truncated with the model's tokenizer. Scoring uses brute-force MaxSim: for our model, the eval-time length-normalised weighted score s_{n} (Eq. 4); for the baselines, the plain MaxSim of Eq. 1. We re-evaluate every public baseline under the same protocol so that all numbers in Table 1 are directly comparable; for Reason-ModernColBERT, where the original release used per-split tuned query lengths, we additionally report the published number with the dagger marker.

OOD evaluation. We additionally evaluate on **OBLIQ-Bench** (Tchuindjo et al., 2026), which probes oblique queries (stance / intent / irony) where similarity-based retrievers are known to fail. We use this as a deliberate OOD stress-test, not as a target.

Baselines. (i) BM25 (Robertson and Zaragoza, 2009), with split-level numbers taken from the BRIGHT release. (ii) `mxbai-edge-colbert-v0-32m` (Takehi et al., 2025) (32M), the upstream public general-purpose ColBERT and the starting point of our recipe. (iii) `mxbai-edge-colbert-v0-17m` (Takehi et al., 2025) (17M). (iv) `answerai-colbert-small-v1` (Clavié, 2024) (33M, BERT-based; document length capped at 512 to respect its position-embedding limit). (v) `GTE-ModernColBERT-v1` (Chaffin, 2025a) (150M), the same backbone family as Reason-ModernColBERT but trained on general-purpose data. (vi) `Reason-ModernColBERT` (Chaffin, 2025b) (150M), built on GTE-ModernColBERT-v1 and fine-tuned on ReasonIR data, our direct quality reference.

Hardware. Base training takes about 24 hours on 8$\times$H100 across two nodes; head training takes about twelve minutes on a single H100.

Table 2: **Classical-IR on NanoBEIR.** Mean nDCG@10 over the NanoBEIR suite; the *No Touche* column drops the NanoTouche-2020 split, which is an outlier across all three retrievers.

Model	All	No Touche
mxbai-edge-32m	60.47	65.51
SmallReason, no head	60.93	65.35
SmallReason	60.00	65.00

Table 3: OBLIQ-Bench out-of-distribution evaluation: Gold nDCG@10 ($\times 100$). We list a representative selection of single-stage retrievers and a lightweight agentic baseline from (Tchuindjo et al., 2026).

Model	Twit.	Wild.	Math	Writ.	Cong.	Mean
Ours (32M)	0.03	0.45	8.54	6.85	2.42	3.66
Qwen3-Emb-4B	3.20	3.10	9.50	3.30	4.00	4.62
Qwen3-Emb-0.6B	0.80	5.90	11.60	4.60	0.60	4.70
LateOn (149M)	0.40	0.20	11.20	10.50	8.30	6.12
GPT-5.2 Rewriter	6.60	6.50	14.20	1.80	8.40	7.50

5 Results

5.1 Main BRIGHT Comparison

Table 1 reports the main BRIGHT comparison. SmallReason-ColBERT reaches 21.41 mean nDCG@10, $+3.25$ over the upstream 32M base (18.16) and $+1.80$ over the same model without the importance head (19.61). It approaches the 150M reasoning-tuned Reason-ModernColBERT (22.62, published; 21.97 under our same-protocol re-evaluation) and trails the 150M general-purpose GTE-ModernColBERT by 0.31 nDCG@10 on the mean. Three observations follow.

The reasoning-tuned base alone improves the upstream on the mean, but remains incomplete. The base scores 19.61, above the upstream's 18.16. The reasoning-aligned curriculum gains substantially on the StackExchange splits (Biology $+4.46$, Earth $+1.99$, Economics $+2.60$, Psychology $+2.98$) and on Pony ($+9.87$), but still regresses on some short, content-dense coding and math splits where the upstream was already strong (AoPS $9.80 \rightarrow 4.89$, TheoremQA-questions $12.51 \rightarrow 9.04$). The base alone, therefore, is not the headline; it is a precondition for the head to work.

The head selectively recovers and extends the reasoning splits. Adding the head lifts the mean by $+1.80$, dominated by the same coding and math splits that the base had regressed: LeetCode $+12.58$, AoPS $+5.40$, TheoremQA-questions $+3.96$. The full system ends up at or above the upstream on every split; the largest regression relative to the un-headed base is Pony, whose 32-token queries are too short for a per-token gate to discriminate informative from uninformative tokens. Per-token weighting is exactly the mechanism long reasoning queries need, and precisely the mechanism an unweighted MaxSim cannot exploit on a base retrained on a length-skewed mixture.

32M with a head approaches 150M ColBERTs. The gap to Reason-ModernColBERT is 1.21 nDCG@10 (published) or 0.56 (same-protocol re-evaluation), and the gap to GTE-ModernColBERT is 0.31. On the StackExchange splits, our model is within 1 nDCG of Reason-ModernColBERT on most splits. This suggests that reasoning-aligned data combined with a tiny query-side gate can recover much of the gap to a $4.7\times$ larger ColBERT on this class of queries. Substantially larger reasoning-tuned single-vector models (ReasonIR-8B at 24.38, Qwen3-Embedding at 22.51) maintain an advantage of one to three nDCG@10 points. Closing that remaining gap may require more capacity or a different inductive bias than a frozen 32M base can provide.

Where the gain comes from. The $+1.80$ mean is dominated by three splits: LeetCode ($+12.58$), AoPS ($+5.40$), and TheoremQA-questions ($+3.96$). All three contain long natural-language statements with a small number of operationally distinguishing tokens, such as function names, math operators, and theorem names. Mild regressions appear on Pony (-3.44), whose 32-token queries leave less room for a gate to separate content from context, Robotics (-0.51), where queries are already content-dense, and Economics (-0.26). The head learns a soft preference over query tokens rather than a hard filter; we characterise this behaviour in §7.

5.2 Out-of-Distribution Evaluation

NanoBEIR. A natural concern is that adding a query-side gate trained on reasoning data hurts classical, keyword-driven IR. We evaluate on NanoBEIR (Thakur et al., 2021) (13 BEIR-derived splits sampled to roughly 50 queries and 1k to 5k documents per split). Table 2 reports the mean nDCG@10 of our model with and without the head, alongside the upstream 32M baseline. The head trades -0.93 nDCG@10 on classical IR for

the $+1.80$ on BRIGHT. Excluding the unstable Touche-2020 split, where the upstream and our final model both score 0.00 and the un-headed base scores 7.88 on 50 queries, the head is nearly flat: $65.35 \rightarrow 65.00$ (-0.35). The trade-off is expected. A query-side gate optimised on long reasoning queries does not transfer cleanly to short keyword queries, but the resulting degradation is modest.

OBLIQ-Bench. OBLIQ-Bench (Tchuindjo et al., 2026) probes oblique queries about stance, intent, behavioural failure, writing style, and tip-of-the-tongue retrieval. Similarity-based retrievers are known to struggle in this setting. Our model performs near zero on the descriptive subsets (Twitter 0.03, WildChat 0.45) and remains low on tip-of-the-tongue queries (Congress 2.42). Table 3 reports our model alongside a selection of single-stage retrievers and a lightweight agentic baseline from (Tchuindjo et al., 2026), all of which outperform our model on the mean despite the absolute scores remaining well below the LLM-level upper bound reported in (Tchuindjo et al., 2026). This is consistent with OBLIQ's thesis that oblique queries require LLM-level verification rather than embedding similarity. We include it as a deliberate negative result and do not claim that a frozen ColBERT with a per-token gate can replace an LLM reranker for this type of query.

5.3 Compute

The head adds little inference cost beyond the ColBERT encoder. For each query token (up to $T_q = 256$) it computes one dot product followed by a sigmoid, which is small compared with the surrounding forward pass. The method still inherits the multi-vector storage and scoring costs of late interaction; our efficiency claim is therefore about parameter count and the additional cost of the head, not about replacing single-vector retrieval. The head trains in roughly twelve minutes on a single H100 with the base frozen, compared with about 24 hours on $8\times$H100 to produce the base itself.

6 Analysis

6.1 Training Asymmetry

The most consequential design choice in the recipe is the training score used for the head. We hold every other choice fixed and replace the un-normalised training loss with the symmetric length-normalised one, so the model trains and evaluates against s_{n}. The cross-entropy loss stalls near $\ln 2 \approx 0.693$, gates barely depart from their $\sigma(5) \approx 0.99$ initialisation, and BRIGHT mean nDCG@10 drops by 3.59. Training and evaluating both against s_{u} is rank-invariant for per-query metrics like nDCG, since a constant per-query rescaling cannot reorder a single query's documents. We therefore use s_{u} for the learning signal and s_{n} for length-comparable scores.

Table 4: Effect of head initialisation on BRIGHT mean nDCG@10. The default setting is the released model; other rows report controlled deltas from the same default.

Initialisation	**Mean**	Δ
$b = 5, \mathbf{W} = \mathbf{0}$	**21.41**	0.00
$b = 0, \mathbf{W} = \mathbf{0}$	21.41	0.00
$b = 10$	21.30	-0.11
$b, \mathbf{W} \sim \mathcal{N}(0, 1)$	21.16	-0.25

6.2 Initialisation Analysis

The identity-at-step-zero initialisation ($b = 5$, $\mathbf{W} = \mathbf{0}$) implements a no-op against the un-headed base before training. We compare against three alternatives in Table 4. Setting $b = 0$ starts gates at 0.5 everywhere; the converged result matches $b = 5$ at this seed, suggesting that the near-identity start is useful mainly as a stable initial condition. Setting $b = 10$ saturates the sigmoid and slightly reduces gradient flow through the bias, costing -0.11 nDCG. A fully random Gaussian initialisation gives the worst result of the four. The differences are small but consistent.

6.3 Hyperparameter Robustness

To verify that the headline number is not a cherry-picked hyperparameter point, we sweep an 18-cell grid in lr $\times$ steps $\times$ batch, retraining the head in each cell and re-evaluating on full BRIGHT. Across the sweep, all runs stay within a 0.42 nDCG@10 range and a 0.11 standard deviation. This supports robustness of the recipe once the un-normalised training loss is in place.

6.4 Head Capacity Analysis

A natural objection is that a single linear gate is under-capacitated. We replaced the 1-layer head with a 2-layer MLP Linear$(d, H) \rightarrow$ ReLU $\rightarrow$ Linear$(H, 1)$ for $H \in \{128, 256, 512, 1024\}$. The output layer is zero-initialised with bias 5, and the hidden layer uses Xavier initialisation so that the step-zero behaviour remains near identity. All four

Table 5: Effect of head capacity on BRIGHT mean nDCG@10. The 1-layer head is the released model; larger heads report controlled deltas from the same default.

Head	Params	Mean	Δ
1-layer linear	129	**21.41**	0.00
2-layer MLP, $H = 128$	16,641	20.68	-0.73
2-layer MLP, $H = 256$	33,281	20.77	-0.64
2-layer MLP, $H = 512$	66,561	20.64	-0.77
2-layer MLP, $H = 1024$	133,121	20.35	-1.06

Table 6: Alternative scoring variants on the frozen reasoning-tuned base. All values are BRIGHT mean nDCG@10. The gate-threshold run covers 10 of the 12 splits (earth science and AoPS are missing); restricted to those same 10 splits the main model scores 20.16 and Top-K ($K = 4$) scores 20.37. Mathematical definitions are given in Appendix B.

Variant	Splits	Mean
SmallReason-ColBERT (main)	12/12	21.41
Top-K MaxSim ($K = 4$)	12/12	21.66
Fixed IDF gate	12/12	20.06
Gate threshold ($w_t < 0.95$)	10/12	16.02

MLPs use the same training schedule and base. Table 5 shows that adding capacity does not improve the result. Every MLP underperforms the 1-layer head, and the largest loses 1.06 nDCG@10 despite having three orders of magnitude more parameters. The frozen base's 128-dimensional projection appears to be the bottleneck; a single linear gate already extracts the useful signal, while additional parameters overfit the short fine-tuning.

6.5 Alternative Score Formulations

We compare the learned head against three closely related score variants on the same frozen base. Table 6 summarizes the results. The comparison clarifies the role of the importance head: Top-K pooling is a small orthogonal improvement, fixed IDF is useful but weaker than the learned gate, and hard-thresholding the learned gates loses the soft weighting signal.

Top-K MaxSim. Replacing $\max_j \mathbf{Q}_t \cdot \mathbf{D}_j$ with the mean of the top K values per query token (Clavié et al., 2025) reaches 21.66 at $K = 4$, a $+0.25$ improvement over our recipe at $K = 1$. This modification changes how each query token pools document evidence, while our contribution changes how query tokens are weighted. We therefore treat Top-K pooling as an orthogonal scoring variant rather than the main contribution.

Fixed IDF anchoring. Replacing the learned gate with a fixed weighting $w_t = \mathrm{IDF}(\mathrm{token}_t)$ reaches 20.06, $+0.45$ over the un-headed base but -1.35 below the learned head. The IDF prior is informative on its own; the learned head finds a different and stronger signal, as we analyse in §7.

Gate thresholding. Setting gates below 0.95 to zero at evaluation time reduces performance to 16.02. The head's contribution lies in relative weighting within a query, not in absolute calibration. Many useful query tokens receive moderate weights, so hard pruning removes signal that the weighted sum still needs.

Table 7: Gate statistics on BRIGHT query tokens (12 splits, $\sim$ 200K tokens). Values are computed over $\sigma(\mathbf{W}\mathbf{Q}_t + b)$ for every query token. Distributions are unimodal and consistent across splits.

Statistic	Min	Median	Max
Per-split mean	0.43	0.45	0.47
Per-split std	0.09	0.10	0.11
Token-level minimum	0.31	–	–
Token-level maximum	–	–	1.00

7 Analysis: What Does the Head Learn?

A natural reading of token-level importance gates is that they approximate inverse document frequency, amplifying rare tokens and suppressing common stopwords. We test this hypothesis directly. We compute the gate value w_t for every query token across all 12 BRIGHT splits ($\sim$ 200,000 tokens in total), pair each token with its IDF in the BRIGHT/earth-science corpus (a representative full-corpus index), and measure the rank correlation between the two.

Table 7 summarises the per-split gate distribution. Across all 12 splits the mean gate sits between 0.43 and 0.47 with a standard deviation of about 0.10, and the distributions are unimodal with a long tail towards 1. The head is doing soft modulation, rather than hard selection. The gate rarely zeros out a token, but it routinely doubles the contribution of certain tokens relative to others within the same query. Across the matched (gate, IDF) pairs the Spearman rank correlation is $\rho = -0.02$ ($p = 0.026$), and the Pearson correlation is -0.04 ($p < 10^{-5}$). Both are statistically detectable on $\sim$ 12,000 pairs but practically zero. The head is *not* learning a soft IDF schedule. This finding is consistent with the IDF-anchoring ablation (§6.5):

a fixed IDF weighting reaches 20.06 (+0.45 over the un-headed base), but the learned head reaches 21.41 (+1.80). Two distinct signals coexist; the IDF prior is informative on its own, but the head finds an additional and stronger signal that has near-zero correlation with token rarity. Qualitative examples are provided in Appendix A.

8 Conclusion

We presented **SmallReason-ColBERT**, a 32M late-interaction retriever for reasoning-heavy retrieval. The model is trained in three stages: a varied-length contrastive warmup, a hard-negative distillation polish, and a single-layer per-query-token importance head trained with an un-normalised weighted MaxSim objective. It reaches 21.41 mean nDCG@10 on BRIGHT, within 1.21 nDCG of the $4.7\times$-larger Reason-ModernColBERT and substantially above all $\leq$ 33M ColBERTs we evaluate. In the controlled ablation, replacing the asymmetric training objective with the symmetric normalised score costs 3.59 nDCG@10; the cross-entropy gradient collapses and the head fails to learn. Through analyses of capacity, initialisation, and score variants we show that the minimal head is a design optimum at this scale: adding capacity to the head hurts, fixed IDF is weaker than the learned gate, and hard thresholding removes useful soft-weighting signal. Analysis of the learned gate values shows that the head is not learning soft-IDF; it amplifies tokens that mark query intent rather than tokens that are rare.

Limitations

Scale of the backbone. The recipe is developed and validated at the 32M scale that this paper targets. We do not claim that the same three-step recipe transfers without modification to substantially smaller backbones, to larger ColBERT models, or to non-ColBERT late-interaction families; the choice of widening, learning rates, and training-data mixtures was tuned to this backbone, and other scales may need different settings.

Frozen-base assumption. The importance head is trained on top of a frozen reasoning-tuned base. We did not explore joint fine-tuning of the base and the head, nor whether the head's contribution persists if the base is further updated. Both are natural follow-ups but lie outside the scope of this paper.

Inference cost of late interaction. Although the head itself adds negligible compute, the model still inherits the multi-vector storage and scoring costs of ColBERT-style late interaction. Our efficiency claim is about parameter count and the marginal cost of the head, not about matching the storage or query-time cost of single-vector retrieval.

Reliance on synthetic teacher data. The warmup and polish stages train on ReasonIR-VL, ReasonIR-HQ, and BGE-Reasoner, whose hard negatives were mined with cross-encoders. Any systematic biases in that negative mining or in the synthetic query distribution can propagate into the resulting retriever, and we have not audited the teacher behaviour beyond using the released data as-is.

Acknowledgments

The authors would like to acknowledge the financial support provided by the Austrian Research Agency (FFG) for the project "AI Enabled Sustainability Jurisdiction Demonstrator" (project No. 915229). The computational results presented in this work have been achieved using the MUSICA cluster, part of the Austrian Scientific Computing (ASC) infrastructure.

References

Abdelrahman Abdallah, Jamie Holdcroft, Mohammed Ali, and Adam Jatowt. 2026a. Are llm-based retrievers worth their cost? an empirical study of efficiency, robustness, and reasoning overhead. *arXiv preprint arXiv:2604.03676*.

Abdelrahman Abdallah, Mohamed Darwish Mounis, Mahmoud Abdalla, Mahmoud SalahEldin Kasem, Mostafa Farouk Senussi, Mohamed Mahmoud, Mohammed Ali, Adam Jatowt, and Hyun Soo Kang. 2026b. Mm-bright: A multi-task multimodal benchmark for reasoning-intensive retrieval. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 8604–8612.

Abdelrahman Abdallah, Bhawna Piryani, Jamshid Mozafari, Andreas Herzinger, Jamie Holdcroft, and Adam Jatowt. 2026c. Rankify: A comprehensive python toolkit for retrieval, re-ranking, and retrieval-augmented generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 208–219.

Mohammed Ali, Abdelrahman Abdallah, Amit Agarwal, Hitesh Laxmichand Patel, and Adam Jatowt. 2026. Recor: Reasoning-focused multi-turn conversational retrieval benchmark. In *Findings of the Association*

for Computational Linguistics: ACL 2026, pages 2688–2723.

Antoine Chaffin. 2025a. Gte-moderncolbert.

Antoine Chaffin. 2025b. Reason-ModernColBERT.

Jianlyu Chen, Junwei Lan, Chaofan Li, Defu Lian, and Zheng Liu. 2025. Reasonembed: Enhanced text embeddings for reasoning-intensive document retrieval. *arXiv preprint arXiv:2510.08252*.

Benjamin Clavié. 2024. Small but mighty: Introducing answerai-colbert-small.

Benjamin Clavié. 2025. Jacolbertv2. 5: Optimising multi-vector retrievers to create state-of-the-art japanese retrievers with constrained resources.

Benjamin Clavié, Sean Lee, Rikiya Takehi, Aamir Shakir, and Makoto P Kato. 2025. Simple projection variants improve colbert performance.

W Bruce Croft, Donald Metzler, Trevor Strohman, and 1 others. 2010. *Search engines: Information retrieval in practice*, volume 520. Addison-Wesley Reading.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning.

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

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3030–3042.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519.

Omar Khattab and Matei Zaharia. 2020. 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*, pages 39–48.

Jimmy Lin and Xueguang Ma. 2021. A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. 4.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. 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*, pages 3715–3734.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, and 1 others. 2025. Reasonir: Training retrievers for reasoning tasks.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Liu Haisu, Quan Shi, Zachary Siegel, Michael Tang, and 1 others. 2025. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval.

Rikiya Takehi, Benjamin Clavié, Sean Lee, and Aamir Shakir. 2025. Fantastic (small) retrievers and how to train them: mxbai-edge-colbert-v0 tech report.

Diane Tchuindjo, Devavrat Shah, and Omar Khattab. 2026. Obliq-bench: Exposing overlooked bottlenecks in modern retrievers with latent and implicit queries.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. 2025. Seq vs seq: An open suite of paired encoders and decoders.

A Qualitative Gate Examples

Inspecting the per-token gates on individual BRIGHT queries clarifies what the head amplifies. On natural-language earth-science queries, the gate is a narrow soft modulation around its per-split mean of 0.43–0.47, with a small saturated tail. On a query about Hadley cells in Jupiter's atmosphere, for example, the gate places "to" and "What" at 1.00, while the content tokens "Hadley" (0.49), "Jupiter" (0.52) and "cells" (0.41–0.56) sit near the mean. Saturation is rare and is not exclusively tied to intent markers: across the five sampled queries

only three tokens exceed 0.90, and one of them is the content noun "map"; sentence-final "?" ranges from 0.43 to 0.68. The head therefore acts as a soft re-weighting over query tokens rather than a selector, and the tokens it amplifies most are not the rare content words an IDF prior would favour — consistent with the near-zero gate–IDF correlation in §7.

B Scoring Functions

This appendix spells out the scoring functions used in the main model and in Table 6. Let

$$M_t(q,d) = \max_{j\in\{1,\dots,T_d\}} \mathbf{Q}_t \cdot \mathbf{D}_j$$

be the standard ColBERT MaxSim value for query token t against document d.

Plain MaxSim. The un-headed base uses the standard additive ColBERT score:

$$s_{\text{plain}}(q,d) = \sum_{t=1}^{T_q} M_t(q,d). \tag{5}$$

This is the path used by the plain baseline evaluator: after computing the query-token-by-document-token similarity matrix, the implementation takes a maximum over document tokens and sums over query tokens.

Learned weighted MaxSim. SmallReason-ColBERT adds a query-side scalar gate

$$w_t = \sigma(\mathbf{W}\mathbf{Q}_t + b), \tag{6}$$

where the one-layer head has $\mathbf{W} \in \mathbb{R}^{1\times 128}$ and $b \in \mathbb{R}$. The base encoder is frozen, so gradients update only $\mathbf{W}$ and b.

During head training, the code uses the un-normalised weighted score:

$$s_{\text{train}}(q,d) = \sum_{t=1}^{T_q} w_t M_t(q,d). \tag{7}$$

At evaluation time, the code first normalises the weights within each query,

$$\bar{w}_t = \frac{w_t}{\sum_{r=1}^{T_q} w_r + \epsilon}, \tag{8}$$

and then scores

$$s_{\text{eval}}(q,d) = \sum_{t=1}^{T_q} \bar{w}_t M_t(q,d) = \frac{\sum_t w_t M_t(q,d)}{\sum_t w_t + \epsilon}. \tag{9}$$

The small ϵ is an implementation guard against division by zero. The normalisation factor is constant across candidate documents for a fixed query, so it does not change per-query rankings, but it keeps scores on a comparable scale across queries.

Top-K MaxSim. The Top-K variant replaces the single best document-token match with the average of the top K matches for each query token. Let $\mathrm{TopK}_K(\mathbf{Q}_t\mathbf{D}^\top)$ denote the set of the K largest token similarities for query token t. The per-token score is

$$M_t^{(K)}(q,d) = \frac{1}{K} \sum_{z\in \mathrm{TopK}_K(\mathbf{Q}_t\mathbf{D}^\top)} z, \tag{10}$$

and the corresponding weighted score is

$$s_{\mathrm{top}K}(q,d) = \sum_{t=1}^{T_q} \bar{w}_t M_t^{(K)}(q,d). \tag{11}$$

Table 6 reports the best value we observed, $K = 4$. This variant is orthogonal to the learned importance head: it changes how each query token pools evidence from document tokens, not how query tokens are weighted.

Fixed IDF gate. The fixed-IDF variant replaces the learned gate with a non-parametric token rarity weight. For a token x_t, let

$$\mathrm{IDF}(x_t) = \log \frac{N}{\mathrm{df}(x_t)}, \tag{12}$$

where N is the number of documents in the reference corpus and $\mathrm{df}(x_t)$ is the number of documents containing the token. The score is

$$s_{\mathrm{IDF}}(q,d) = \frac{\sum_t \mathrm{IDF}(x_t) M_t(q,d)}{\sum_t \mathrm{IDF}(x_t) + \epsilon}. \tag{13}$$

This tests whether the learned head is merely recovering a lexical rarity prior. It improves over the un-headed base, but remains below the learned head.

Gate thresholding. The thresholding variant keeps the learned gate values but turns them into a hard mask:

$$\tilde{w}_t = \begin{cases} w_t, & w_t \geq 0.95, \\ 0, & w_t < 0.95. \end{cases} \tag{14}$$

The score then uses the same normalised weighted form,

$$s_{\mathrm{thr}}(q,d) = \frac{\sum_t \tilde{w}_t M_t(q,d)}{\sum_t \tilde{w}_t + \epsilon}. \tag{15}$$

Its regression in Table 6 shows that the head should be interpreted as a soft weighting mechanism rather than as a calibrated token selector.