

Depth-Selective Residual Routing for Dense Retrieval: Block Attention Residual Adapters under Cross-Domain and Quantization Controls

CHELATEDAI Research Archive

Independent research field note series

PUBLISHED 26 AUGUST 2026 6,188 WORDS 15 REFERENCES ADAPTIVE-RETRIEVAL

Evidence boundary. implemented surface with negative promotion evidence. This manuscript is a public preprint and is not represented as peer reviewed or as proof of production advantage.

CHELATEDAI Research Archive

Author record: CHELATEDAI Research Archive

Manuscript type: Methods and negative-results paper

Version: 1.0, 26 August 2026

Abstract

Dense retrievers commonly expose a fixed-dimensional embedding interface: a query and each document are encoded, optionally transformed by a compact adapter, normalized, and compared by cosine similarity. A conventional residual adapter makes one correction and commits to it. This paper studies a different interface, **BlockAttnResAdapter**, in which a sequence of residual correction blocks produces candidate states at multiple correction depths and a learned attention operator selects a convex mixture of those states for each input. The design is adapted from Attention Residuals, which replaces uniform residual accumulation across Transformer depth with learned depth-wise aggregation, but the object studied here is narrower: a post-encoder retrieval adapter rather than a pretrained Transformer residual stream.

We report the complete public CHELATEDAI implementation and evaluation record available for this adapter. The implementation includes two-, four-, and eight-block profiles; near-identity initialization; L2-normalized output; an adapter factory; bounded-wrapper compatibility; a companion cross-layer aggregator; deterministic road-course profiles; and a machine-checked promotion summary. Forty-five focused adapter tests and one decision-summary test pass in the current source checkout. The empirical record consists of six

documented small-slice runs using `sentence-transformers/all-MiniLM-L6-v2`, 20 queries, 1,200 sampled documents, two training epochs for trained profiles, SciFact and NFCorpus, and seeds 42, 43, and 44. Balanced BlockAttnRes is negative or tied on all three SciFact seeds. It is positive relative to the untrained baseline on all three NFCorpus seeds, with NDCG@10 deltas of +0.020078, +0.001119, and +0.001117, but only one of six run-level quantization-survival gates passes. The repository's aggregate decision is therefore `no_default_change`.

These results establish implementation viability and reveal task-conditioned behavior; they do not establish general retrieval efficacy. Raw run artifacts referenced by the committed summaries are not present in the reviewed checkout, precluding independent metric recomputation from ranked lists and qrels. The sampled test sets are small, the training target is shared with the MLP comparator, and there is no untouched hyperparameter-selection set, full-corpus evaluation, confidence interval, or cross-encoder control. We consequently present a falsifiable next-stage protocol centered on immutable qrels, per-query paired statistics, matched parameter and compute budgets, explicit depth-selection diagnostics, and a quantization-aware veto. The contribution is a precise residual-routing hypothesis, a reproducible implementation record, and a non-promotion boundary—not a claim of a superior retrieval adapter.

Keywords: dense retrieval; residual adapters; attention residuals; depth-wise routing; parameter-efficient adaptation; BEIR; quantization; negative results; reproducibility

1. Introduction

Bi-encoder retrieval separates representation from search. An encoder maps queries and documents to fixed-width vectors; a similarity function then ranks a corpus without jointly processing every query-document pair. Sentence-BERT established the practical form of this architecture for semantic search by training embeddings that can be compared directly with cosine similarity [1]. The separation is computationally valuable, but it also creates a narrow intervention point. A post-encoder transformation can improve, preserve, or damage the geometry used for ranking while leaving the underlying language model unchanged.

Compact adapters are one family of such interventions. In parameter-efficient transfer learning, adapters add a small number of trainable parameters while keeping a pretrained backbone fixed [2]. Low-Rank Adaptation (LoRA) similarly constrains trainable updates to low-rank factors [3]. Those methods differ in placement and parameterization, but both motivate a useful engineering principle: isolate adaptation from the base model so that task-specific behavior is bounded, versionable, and removable. In retrieval, a lightweight adapter can be trained to map embeddings toward a new corpus, encoder version, or supervision signal. The most common form is a single residual map,

$$[\hat{y} = \operatorname{norm}(\mathbf{1} + f_{\theta}(x)),]$$

where $x \in \mathbb{R}^d$, f_{θ} is a small network, and normalization returns the vector to the cosine-retrieval sphere. This design is cheap and interpretable as a correction. It has no internal mechanism for selecting how much correction depth an individual input requires. If the learned direction is unhelpful, the model can only reduce the residual magnitude globally or be retrained.

Attention Residuals (AttnRes) offers a neighboring idea. Kimi Team et al. replace the uniform accumulation of preceding layer outputs in deep PreNorm Transformers with learned, input-dependent attention across depth [4]. Their Block AttnRes variant groups layers and attends over block-level representations, reducing storage and communication relative to full depth-wise attention. The original result concerns residual streams during large-scale model pretraining. It does not imply that the same mechanism improves frozen dense embeddings. Nevertheless, its central operation—choosing among intermediate representational states—can be translated into a post-encoder adapter.

CHELATEDAI implements that translation as `BlockAttnResAdapter`. Starting from the original embedding, it applies a sequence of residual multilayer perceptrons, retains every intermediate state, and uses the terminal state as a query over all retained states. The output is a learned convex combination of correction depths, followed by L2 normalization. A second module, `LayerAttentionAggregator`, applies a related operation to mean-pooled hidden states captured from multiple Transformer layers. These are distinct surfaces: the former transforms one already-produced embedding; the latter aggregates a stack of layer embeddings. The public benchmark record evaluates the former.

The scientific question is not whether the architecture runs. It does. The question is whether depth-selective correction improves retrieval relative to simpler, matched controls and whether any gain survives the representation changes imposed by a deployment path such as quantization. That question requires restraint because retrieval behavior can vary sharply by corpus and relevance structure. BEIR was created precisely because narrow, homogeneous evaluation obscures out-of-domain behavior [5], while MTEB shows that no single embedding method dominates every embedding task [6]. A positive result on one sampled task cannot support a general default.

This manuscript contributes four things.

1. It gives a formal account of the implemented `BlockAttnRes` retrieval adapter, including the distinction between convex depth selection and ordinary residual accumulation.
2. It documents the exact public implementation, focused tests, benchmark profiles, and promotion logic reviewed at a fixed source commit.
3. It reports the complete six-run record rather than the best slice: three non-positive SciFact results, three positive NFCorpus results, and five quantization-gate failures.
4. It specifies a preregistrable protocol capable of confirming or falsifying the claim that block attention residuals add value beyond parameter count, training, and ordinary nonlinear residual depth.

Our conclusion is deliberately limited. The adapter is a valid experimental object. Its observed behavior is task-conditioned. The available evidence does not support changing the default retrieval adapter.

2. Related Work

2.1 Residual streams and depth-wise aggregation

Residual networks introduced identity shortcuts that let a transformation learn a residual relative to its input [7]. Transformers use residual connections around attention and feed-forward sublayers [8]. Modern PreNorm architectures place normalization before a sublayer, which improves optimization stability but causes the residual stream to accumulate outputs across depth. Analyses of normalization placement have shown that PreNorm and PostNorm differ in gradient behavior and training stability [9].

Attention Residuals changes the aggregation rule itself. Rather than giving every preceding contribution an implicit unit coefficient, it assigns learned, input-dependent weights over prior representations [4]. Full AttnRes stores layer-level states; Block AttnRes groups layers so the cross-depth operator acts over fewer block summaries. The claimed benefit in the source paper is improved depth utilization in models trained with the mechanism from scratch. CHELATEDAI borrows the selection principle, not the evidentiary result. A post-hoc embedding adapter operates after the pretrained encoder, at far smaller depth, under different supervision, and without co-adaptation of the base model. Transfer of mechanism is therefore a hypothesis.

Deep cross-layer aggregation is also related to learned mixtures of hidden layers. ELMo learned task-specific scalar mixtures over internal language-model layers [10]. BERT analyses and feature-extraction practices likewise treat layer choice as task dependent. The CHELATEDAI `LayerAttentionAggregator` is closest to this family because it weights captured layer embeddings. `BlockAttnResAdapter`, by contrast, creates its own correction-depth states after encoding. Conflating these modules would overstate the benchmark: the reported road course tests adapter depth, not raw Transformer-layer selection.

2.2 Parameter-efficient adaptation

Houlsby et al. introduced compact bottleneck adapters for NLP transfer while freezing the original network [2]. LoRA freezes pretrained weights and learns low-rank update matrices [3]. Subsequent modular adapter libraries unify multiple strategies behind common composition interfaces [11]. The shared motivation is parameter and operational isolation, not a guarantee that any particular adapter improves every task.

The present adapter is not LoRA. Each correction block contains a down-projection from (d) to $(d/2)$, ReLU, and an up-projection to (d) , and its attention projections are dense. For (B) blocks and projection dimension (p) , ignoring biases, the trainable parameter count is approximately

$$[P_{\{\text{BAR}\}} \approx B \left(d \frac{d}{2} + \frac{d}{2} d \right) + 2dp = Bd^2 + 2dp.]$$

Including biases adds $(B(3d/2) + 2p)$. This can substantially exceed a single bottleneck adapter. Any future comparison must match either trainable parameters, floating-point operations, wall time, or all three; otherwise a depth-selective advantage could be a capacity advantage.

2.3 Dense retrieval evaluation

Sentence-BERT supplies the basic cosine-retrieval setting [1]. BEIR aggregates heterogeneous retrieval datasets and shows that performance rankings differ across domains and model families [5]. MTEB broadens embedding evaluation across tasks, datasets, and languages and similarly finds no universal winner [6]. These findings are directly relevant to the divergent SciFact and NFCorpus observations reported here.

SciFact contains expert-written scientific claims, evidence-bearing abstracts, and rationale annotations [12]. NFCorpus is a medical information-retrieval collection assembled from natural-language queries and full-text relevance links [13]. Their relevance distributions, language, and numbers of relevant documents differ. A mechanism can plausibly respond differently to those structures, but the current experiment does not identify the causal feature responsible for the difference.

Normalized Discounted Cumulative Gain (NDCG) discounts relevant documents by rank and normalizes against the ideal ranking [14]. Mean Average Precision, reciprocal rank, and recall expose complementary behavior. Reporting all four is useful, but metrics computed on a small sampled corpus remain conditional on the sampling and qrels contract. Statistical inference should be based on per-query paired differences, not on treating seeds or aggregate metric rows as independent observations.

2.4 Quantization as a promotion boundary

Quantization reduces storage and arithmetic cost by representing weights or activations at lower precision. Modern methods such as SmoothQuant address activation outliers so that INT8 deployment can preserve accuracy in large language models [15]. The CHELATEDAI gate is narrower: it asks whether a retrieval gain observed in a floating-point candidate survives the repository's quantized path. It is a deployment-oriented veto, not an evaluation of SmoothQuant and not evidence about general quantization methods. The repeated failure is scientifically material because an adapter whose small ranking gain vanishes under the intended representation path is not a robust default candidate.

3. Method

3.1 Baseline retrieval

Let (E) be a frozen text encoder and $(x=E(t)\in\mathbb{R}^d)$ the embedding of text (t) . Given a query (q) and document (d_j) , an adapter (A_{θ}) produces unit vectors

$$[u_q=A_{\theta}(E(q)), \quad u_j=A_{\theta}(E(d_j)),]$$

and scores them by

$$[s(q,d_j)=u_q^{\text{top}}u_j.]$$

Because outputs are L2-normalized, the dot product equals cosine similarity. A baseline profile uses the repository's MLP adapter without training in the benchmark slice. Trained MLP and trained BlockAttnRes profiles share the same two-epoch sedimentation procedure in the reported grid.

3.2 Block attention residual adapter

Let (B) be the number of correction blocks. The adapter initializes a state sequence with the encoder embedding:

$$[h_0 = x.]$$

For block $(b \in \{1, \dots, B\})$, define

$$[f_b(h) = W^{(b)}_{-2} \cdot \sigma(W^{(b)}_{-1}h + b^{(b)}_{-1}) + b^{(b)}_{-2},]$$

where $(W^{(b)}_{-1} \in \mathbb{R}^{d/2 \times d})$, $(W^{(b)}_{-2} \in \mathbb{R}^{d \times d/2})$, and (σ) is ReLU. States advance through ordinary residual updates:

$$[h_b = h_{b-1} + f_b(h_{b-1}).]$$

All states $(H = [h_0, h_1, \dots, h_B])$ are retained. The terminal state supplies a query, while each retained state supplies a key:

$$[q = W_q h_B + b_q, \quad k_b = W_k h_b + b_k,]$$

with $(W_q, W_k \in \mathbb{R}^{p \times d})$. The attention logits and weights are

$$[\ell_b = \frac{q^{\top} k_b}{\sqrt{p}}, \quad \alpha_b = \frac{\exp(\ell_b)}{\sum_{r=0}^B \exp(\ell_r)}.]$$

The output before normalization is

$$[z = \sum_{b=0}^B \alpha_b h_b,]$$

and the retrieval vector is

$$[A_{\theta}(x) = \frac{z}{\|z\|_2}.]$$

The softmax gives $(\alpha_b \geq 0)$ and $(\sum_b \alpha_b = 1)$, so (z) lies in the convex hull of the intermediate states. This bounds the aggregation in a geometric sense, but it does **not** bound the states themselves: repeated residual blocks can still increase their norms. Nor does convex mixing guarantee ranking stability after normalization. Those are empirical properties to measure, not consequences of the formula.

Near-identity initialization uses Gaussian weights with standard deviation 0.001 and zero biases for every correction and attention projection. With tiny correction weights, successive (h_b) initially remain close to (x) . Small query and key projections also make logits similar, so initial attention is approximately uniform. The adapter therefore begins near a normalized average of nearly identical states. “Near identity” is accurate; exact identity is not, because finite random weights and final normalization alter the vector.

The default projection dimension is $(p = \max(\lfloor d/4 \rfloor, 32))$. The public presets select $(B \in \{2, 4, 8\})$, named shallow, balanced, and deep. The factory exposes the type as `attnres`, and the implementation accepts one- or two-dimensional input tensors, preserves the original rank, normalizes outputs, validates save/load paths, and can be wrapped by the repository's bounded-adapter interface [R1].

3.3 Companion layer-attention aggregator

For completeness, the repository includes a separate `LayerAttentionAggregator`. Given mean-pooled Transformer layer embeddings $(G = [g_1, \dots, g_L] \in \mathbb{R}^{L \times d})$, it forms a neutral query from their mean,

$$[\bar{g} = \frac{1}{L} \sum_{\ell=1}^L g_{\ell}, \quad q_g = W_q \bar{g},]$$

computes keys $(k_{\ell} = W_k g_{\ell})$, attends over layers, and normalizes the weighted result. This module has an implemented runtime bridge and focused shape/gradient guards, but the six-run road-course evidence does not evaluate its utility. Any claim about learned Transformer-layer aggregation remains unrun in the record reviewed here.

3.4 Hypotheses

The architecture implies testable, separable hypotheses.

- **H1, depth selection:** On at least one preregistered retrieval distribution, BlockAttnRes improves held-out per-query utility over a parameter- and compute-matched residual MLP.
- **H2, selective mechanism:** The learned attention weights vary with inputs and this variation predicts which correction depth yields the best counterfactual ranking. A constant or uniformly averaged mixture should not explain the gain.
- **H3, transfer:** Any benefit persists across seeds and at least one untouched corpus rather than arising only on the selection task.
- **H4, deployability:** The benefit survives the declared quantized representation path within a preregistered retention margin.
- **H5, efficiency:** Quality-per-latency or quality-per-trainable-parameter is better than the matched control.

The existing experiments address fragments of H1, H3, and H4. They do not directly test H2 or H5.

4. Implementation and Experimental Record

4.1 Exact implementation surface

The reviewed public source is CHELATEDAI commit `65c9085cd048e8a7351a53e87666fdd5639e612b`. The relevant tracked files are unchanged in the local working tree relative to that commit. `chelation_adapter.py` defines `BlockAttnResAdapter`, `LayerAttentionAggregator`, and factory integration [R1]. `config.py` defines the default four-block configuration and two-, four-, and eight-block presets [R2].

`run_road_course_campaign.py` defines the comparison grids and temporarily binds the selected adapter type, block count, and projection dimension while constructing an engine [R3].

`attnres_repeat_seed_decision.py` loads six declared JSON artifacts and fails promotion if any delta is non-positive or any quantization gate fails [R4].

The current checkout passes:

- `python -m unittest test_attnres_adapter` : 45 tests;
- `python -m unittest test_attnres_repeat_seed_decision` : 1 test.

The adapter suite covers shapes, normalization, near-identity behavior, block counts, gradient propagation, factory construction, bounded wrapping, persistence, configuration copies, invalid inputs, and the layer aggregator. Passing tests establish the coded contracts under their fixtures. They do not establish retrieval benefit.

4.2 Initial untrained comparison

The first documented road-course run used SciFact, seed 42, 20 queries, 1,200 sampled documents, and `sentence-transformers/all-MiniLM-L6-v2`. An untrained balanced AttnRes path tied the MLP baseline at NDCG@10 0.816574, MAP@10 0.771726, MRR 0.770833, and Recall@10 0.950000 [R5]. A tied near-identity run verifies that the new route can execute without visible regression on that slice; it is not positive efficacy evidence.

4.3 Trained depth grid

The trained grid used two sedimentation epochs after an opt-in centering warmup. On SciFact seed 42, the untrained MLP baseline obtained NDCG@10 0.816574. The trained MLP and all trained AttnRes depths—two, four, and eight blocks—obtained 0.798669 [R6]. Their identical aggregate scores indicate that the current training path, not depth selection alone, is a plausible source of damage. It does not show that the models learned identical functions.

On NFCorpus seed 42, the baseline obtained NDCG@10 0.837911, the trained MLP 0.853883, the two-block profile 0.853883, the four-block profile 0.857989, and the eight-block profile 0.851135 [R7]. The four-block candidate led the local grid by +0.020078 over baseline. Its quantized evaluation obtained 0.837457, below

the baseline by 0.000453; the retained-gain ratio was negative and the gate failed. Thus the float-path signal and deployment-path signal disagree.

4.4 Repeat-seed record

The repository documents seeds 42, 43, and 44 for SciFact and NFCorpus. The balanced candidate results are reproduced below [R8, R9]. Values are descriptive aggregates from committed reports, not recomputations performed for this manuscript.

Task	Seed	Baseline NDCG@10	Trained MLP	Balanced AttnRes	Candidate delta vs. baseline	Quantization gate
SciFact	42	0.816574	0.798669	0.798669	-0.017905	fail
SciFact	43	0.842134	0.823680	0.817134	-0.025000	fail
SciFact	44	0.840797	0.840797	0.840797	0.000000	fail
NFCorpus	42	0.837911	0.853883	0.857989	+0.020078	fail
NFCorpus	43	0.863763	0.863362	0.864882	+0.001119	fail
NFCorpus	44	0.874058	0.873912	0.875174	+0.001117	pass

Across these six observations, balanced AttnRes is negative twice, tied once, and positive three times. All positive observations occur on NFCorpus. The descriptive mean candidate delta is approximately -0.014302 on SciFact and +0.007438 on NFCorpus; pooling the six heterogeneous observations gives approximately -0.003432. These means are not inferential estimates: three seeds over one fixed 20-query slice do not justify a population claim, and averaging across tasks can conceal precisely the interaction of interest.

The committed decision summary uses the best profile in each run to classify three positive NFCorpus rows and three non-positive SciFact rows. Five of six quantization gates fail. Its machine-readable conclusion is `no_default_change`, with reasons `repeat_seed_non_positive_delta_present`, `quantization_survival_failures_present`, and `task_seed_consistency_failed` [R9].

4.5 Evidence boundary

The committed documents name six raw artifacts under `experiment_runs/roadcourse-small/`. Those artifacts are not present in the reviewed checkout. Consequently, this manuscript cannot independently reconstruct sampled query identifiers, candidate identifiers, per-query ranks, full qrels, attention weights, training trajectories, or metric calculations. The tables above are a faithful synthesis of tracked Markdown and JSON summaries, not an independent audit of the underlying rankings.

This missing layer changes the strength of the evidence. The record supports the statement “the public repository reports these bounded runs and encodes a non-promotion decision.” It does not support a stronger claim that an independent reader can currently reproduce every reported decimal from immutable artifacts. Publication-grade regeneration should remedy this.

5. Analysis

5.1 What the results do establish

First, the architecture is operational. It constructs multiple residual states, routes gradients through blocks and attention projections, integrates with the adapter factory and benchmark grid, and can be trained under the same harness as the MLP comparator. Second, near-identity initialization prevents an immediate regression on the first untrained SciFact comparison. Third, the training and task interaction is nontrivial: SciFact degrades or ties while NFCorpus improves slightly in all three documented seeds. Fourth, a float-path advantage is insufficient under the project's declared acceptance rule because most gains do not survive quantization.

These are useful results. Negative transfer identifies a boundary that a generic “more adaptive” story would miss. The NFCorpus pattern identifies a candidate interaction worth studying. The quantization result identifies a deployment mismatch early, before a default is changed.

5.2 What the results do not establish

The current record cannot attribute the NFCorpus difference specifically to content-dependent attention. Balanced AttnRes has different capacity and optimization geometry from the MLP. No uniform-mixture ablation removes learned attention while preserving the correction blocks. No last-state-only ablation tests whether depth itself, rather than selection, is responsible. No shuffled-attention or frozen-attention control tests whether input dependence matters. The repository summaries do not report attention entropy, selected depth, state norms, or per-query counterfactual best depth.

The experiment also does not establish statistical reliability. Twenty queries are vulnerable to a few rank changes. Seeds appear to change sampling and training together, complicating causal interpretation. The same datasets are used to observe the candidate and discuss next steps; there is no declared development/test split at the experiment-family level. Latency differences of roughly one or two milliseconds are reported, but no repeated timing distribution, warmup protocol, device specification, or uncertainty is available.

Finally, the quantization gate exposes fragility but does not diagnose it. The loss could come from quantizing adapter outputs, thresholds, normalization interactions, or another path in the harness. Without saved float and quantized vectors, neighbor-overlap statistics, and per-query rank changes, “quantization erased the gain” is the endpoint, not the mechanism.

5.3 Interpreting the task interaction

Several explanations remain compatible with the evidence. NFCorpus may contain query or relevance structures that benefit from a modest nonlinear transformation shared by both trained MLP and AttnRes; indeed, the trained MLP is already positive on seed 42. The additional four-block mixture may help a subset of queries, but the +0.0011 deltas on seeds 43 and 44 are small enough to require per-query inspection and uncertainty estimates. SciFact may be more sensitive to perturbing a strong base geometry, or the sedimentation target may be poorly matched to its relevance structure. Different numbers of relevant documents per query can also make NDCG and recall respond differently.

The correct interpretation is an interaction hypothesis: **under the current sampling and training procedure, adapter training behaves differently on SciFact and NFCorpus, and balanced AttnRes is the best local candidate on NFCorpus.** This is narrower than “BlockAttnRes improves biomedical retrieval” and much narrower than “BlockAttnRes improves retrieval.”

6. Falsifiable Confirmation Protocol

6.1 Freeze before tuning

A confirmatory campaign should publish an immutable protocol containing repository commit, environment lock, encoder revision, dataset revisions, qrels hashes, query inclusion rules, document sampling algorithm, random seeds, metric implementation, tie policy, hardware class, quantization transform, and stopping rules. Candidate selection should use a development partition or designated datasets. At least one evaluation corpus and its seeds must remain untouched until architecture, loss, and gates are frozen.

The raw artifact for every run should include query and document IDs, original text hashes where licensing permits, complete qrels used by the metric, ranked lists, raw scores, float and quantized embeddings or stable hashes plus a verifier, attention weights, intermediate-state norms, training events, timing samples, and configuration. A deterministic verifier should recompute aggregate metrics and the promotion decision.

6.2 Matched model family

The minimum model grid is:

1. frozen encoder with identity normalization only;
2. current MLP adapter;
3. deeper residual MLP with the same number of correction blocks but last-state output;
4. same blocks with a uniform average over states;
5. same blocks with learned global, input-independent mixture weights;
6. BlockAttnRes with input-dependent attention;
7. BlockAttnRes with attention weights shuffled across queries at evaluation;

8. two-, four-, and eight-block BlockAttnRes profiles.

Comparisons should report both a parameter-matched grid and a compute-matched grid. If exact matching is impossible, parameter count, multiply-adds, peak memory, training time, and inference latency must be reported so the tradeoff remains visible.

6.3 Training controls

The current sedimentation target should be compared with at least one ordinary contrastive retrieval objective using identical positives, negatives, batches, and training steps. Hyperparameters must be selected without looking at the held-out confirmation set. The baseline MLP and all BlockAttnRes variants should receive matched optimizer, batch, negative-mining, checkpoint-selection, and early-stopping budgets.

Training should include an explicit no-training control, because the initial tie is evidence about safe initialization. It should also include a “training-only” control that applies the same objective to the base adapter family. A proposed attention benefit is credible only if it exceeds the effect of training a simpler adapter.

6.4 Outcomes and statistics

The primary endpoint should be preregistered, for example macro NDCG@10 over all eligible queries. Secondary endpoints may include MAP@10, MRR, Recall@10, success at larger cutoffs, and top-k neighbor overlap. Statistical analysis should use per-query paired differences between the frozen candidate and each matched control. A paired randomization test or stratified bootstrap can provide uncertainty without treating seeds as independent queries. Report effect size, 95% interval, and the fraction of queries improved, tied, and harmed.

Promotion should require all of the following:

- a positive primary-effect interval against the strongest matched control on the untouched confirmation corpus;
- no preregistered catastrophic-harm excess in per-query deltas;
- directionally consistent results across the required tasks and seeds, or a preregistered task router with its own held-out test;
- a quantized-path retained-gain ratio above the frozen threshold and quantized fitness not below baseline;
- acceptable latency, memory, and training-cost ceilings; and
- complete artifact verification from immutable inputs.

Failure of any hard gate returns `no_default_change`. A task-specific positive result may remain a research candidate without becoming a global default.

6.5 Mechanism tests

To test whether attention is doing meaningful depth selection, record $(\alpha(x))$ for every query and document. For each input, also evaluate counterfactual outputs using each individual state (h_b), the uniform mixture, and the learned mixture. Define the oracle depth

$$[b^*(x) = \arg\max_b U(h_b; x),]$$

where (U) is a query-level retrieval utility evaluated without using that query for training. The mechanism hypothesis predicts that attention mass should correlate with useful depths and that the learned mixture should outperform uniform and final-state controls. Attention entropy

$$[H(\alpha) = -\sum_{b=0}^B \alpha_b \log \alpha_b]$$

should be analyzed against query difficulty and gain. Uniform high entropy with no counterfactual advantage would suggest that extra depth, not selection, explains behavior. Extremely low entropy selecting the same block for every input would suggest an ordinary depth choice could replace attention.

State norms and angular movement should also be reported:

$$[r_b(x) = \frac{\|h_b - h_{b-1}\|_2}{\|h_b\|_2 + \epsilon}, \quad \theta_b(x) = \arccos \frac{h_b^{\top} h_{b-1}}{\|h_b\|_2 \|h_{b-1}\|_2}.]$$

These diagnostics separate harmless rescaling from rank-relevant rotation and expose whether deeper profiles accumulate unstable corrections.

6.6 Quantization diagnosis

For each candidate, preserve paired float and quantized outputs. Report cosine error, norm error before final normalization, top-k overlap, rank-biased overlap, and per-query metric delta. If a gain disappears, determine whether error concentrates in attention logits, intermediate states, final vectors, or the search index. A quantization-aware training variant is justified only after this localization; otherwise it risks optimizing against an unspecified failure.

The falsifier is straightforward. If input-dependent BlockAttnRes fails to exceed last-state, uniform-mixture, and global-mixture controls under matched budgets, H2 is rejected even if the larger model beats a shallow MLP. If a float gain repeatedly fails the frozen quantization gate, H4 is rejected for that deployment path. If gains remain confined to one tuned corpus, H3 is rejected.

7. Limitations

This study is an archive synthesis of an existing code and benchmark record, not a newly executed full campaign. The current validation reruns focused unit tests but not the retrieval experiments. The committed summary documents do not ship the named raw run artifacts in the reviewed checkout. Independent readers

therefore cannot reconstruct the six metrics from rankings and qrels using this site alone.

The evaluation covers only two tasks, three seeds, and 20 sampled queries per run. It uses one compact encoder. The sample is too small for broad claims and may be sensitive to the deterministic corpus-sampling procedure. There is no evidence for larger encoders, multilingual data, approximate nearest-neighbor indexes, long documents, dynamic corpora, or production traffic.

The MLP control is operationally relevant but not fully capacity matched. The training objective may dominate the architecture effect, as suggested by identical SciFact seed-42 results across trained profiles. The exact semantic content of “sedimentation” must be evaluated against standard contrastive objectives before attributing results to residual attention.

The reported latency values are not benchmark-grade systems measurements. There is no energy measurement, memory profile, repeated timing distribution, or device-normalized throughput. The adapter's dense (Bd^2) block cost can be material despite being small relative to an encoder.

Attention weights are not explanations by default. Even if their distribution correlates with retrieval gain, causal mechanism requires ablations. The present record does not include those weights.

8. Ethics and Broader Impact

SciFact and NFCorpus concern scientific and medical information. Better ranking can help users find evidence, but a small or unstable change can also suppress relevant material. Nothing in this paper validates the system for clinical decision support, scientific fact verification, or any safety-critical use. The adapter operates on relevance geometry; it does not verify truth.

Publishing negative and mixed results has a positive procedural impact. It discourages promotion from a best seed, makes deployment transformation part of the scientific claim, and distinguishes code correctness from model utility. The same rigor can slow apparent progress, but it reduces the risk of shipping an adaptive component whose gains are corpus-specific or precision-fragile.

The datasets and pretrained encoder carry the biases of their source material and training data. A future campaign should stratify outcomes where meaningful, document licensing and privacy constraints, and avoid treating medical relevance labels as patient-level evidence.

9. Availability and Reproducibility

The CHELATEDAI repository is public at github.com/mattmre/CHELATEDAI. This manuscript reviews tracked surfaces at commit `65c9085cd048e8a7351a53e87666fdd5639e612b`. Relevant files are linked in “Repository evidence cited” below.

Focused validation on 26 August 2026 produced 45 passing tests in `test_attnres_adapter` and one passing test in `test_attnres_repeat_seed_decision`. No retrieval campaign was rerun. The six `experiment_runs/roadcourse-small/attnres_*.json` artifacts named by the decision script are absent from the reviewed checkout; regeneration and durable publication of those artifacts are required for full computational reproducibility.

10. Status and Claim Boundary

Publication status: Archive manuscript; not peer reviewed. It is formatted for scholarly review but is not represented as accepted by arXiv, indexed by Google Scholar, or published by a conference or journal.

Confirmed in the current checkout: the adapter, layer aggregator, presets, road-course grids, and fail-closed decision summarizer exist; their 46 focused tests pass; tracked source files are unchanged relative to the cited commit.

Reported by committed experiment summaries: the six small-slice results in Table 1; three positive NFCorpus candidate deltas; no positive SciFact candidate delta; and five quantization-gate failures.

Not independently reproduced here: dataset download, slice selection, training, ranking, raw artifact generation, or metric recomputation.

Not established: general retrieval improvement; causal value of input-dependent depth attention; superiority at matched parameter or compute budgets; robust cross-domain transfer; quantization-safe benefit; full-corpus utility; production latency; or a basis for changing defaults.

11. Conclusion

Block attention residuals define a coherent retrieval-adapter hypothesis. Instead of committing to the last state of a correction network, the adapter retains multiple depths and learns a convex, input-dependent route through them. The implementation is real, bounded, and well covered by focused tests. The experimental record is also informative precisely because it is mixed: the candidate is negative or tied on SciFact, positive on three NFCorpus seeds, and rejected by five of six quantization gates.

That pattern is a research lead, not a win. It could reflect meaningful corpus-conditioned depth selection; it could also reflect ordinary training effects, extra capacity, sampling noise, or a small set of influential queries. The present record lacks the ablations and immutable raw artifacts needed to distinguish these explanations. The correct decision is therefore the repository's existing one: no default change.

A decisive next study must compare learned attention with last-state, uniform, and global-mixture controls; match capacity and compute; freeze an untouched confirmation set; analyze per-query effects; publish complete qrels-bound artifacts; and require the benefit to survive quantization. Under that protocol, BlockAttnRes can earn a narrow claim, fail cleanly, or be reduced to a simpler mechanism. All three outcomes would advance the work.

References

- [1] Nils Reimers and Iryna Gurevych. 2019. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” *Proceedings of EMNLP-IJCNLP 2019*, 3982–3992. DOI: 10.18653/v1/D19-1410. ACL Anthology.
- [2] Neil Houlsby, Andrei Giurgiu, Stanisław Jastrzębski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. “Parameter-Efficient Transfer Learning for NLP.” *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:2790–2799. PMLR. arXiv: 1902.00751.
- [3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. “LoRA: Low-Rank Adaptation of Large Language Models.” *International Conference on Learning Representations*. OpenReview. arXiv: 2106.09685.
- [4] Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, et al. 2026. “Attention Residuals.” arXiv: 2603.15031. Official repository.
- [5] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. “BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models.” *NeurIPS 2021 Datasets and Benchmarks Track*. arXiv: 2104.08663. OpenReview.
- [6] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. “MTEB: Massive Text Embedding Benchmark.” *Proceedings of EACL 2023*, 2014–2037. DOI: 10.18653/v1/2023.eacl-main.148. ACL Anthology.
- [7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. “Deep Residual Learning for Image Recognition.” *Proceedings of CVPR 2016*, 770–778. DOI: 10.1109/CVPR.2016.90. arXiv: 1512.03385.
- [8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *Advances in Neural Information Processing Systems 30*. Proceedings. arXiv: 1706.03762.
- [9] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. 2020. “On Layer Normalization in the Transformer Architecture.” *Proceedings of ICML 2020*, PMLR 119:10524–10533. PMLR. arXiv: 2002.04745.
- [10] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. “Deep Contextualized Word Representations.” *Proceedings of NAACL-HLT 2018*, 2227–2237. DOI: 10.18653/v1/N18-1202. ACL Anthology.
- [11] Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2023. “Adapters: A Unified Library for Parameter-Efficient and Modular Transfer Learning.” arXiv: 2311.11077.

- [12] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. “Fact or Fiction: Verifying Scientific Claims.” *Proceedings of EMNLP 2020*, 7534–7550. DOI: 10.18653/v1/2020.emnlp-main.609. ACL Anthology.
- [13] Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. “A Full-Text Learning to Rank Dataset for Medical Information Retrieval.” *Advances in Information Retrieval: ECIR 2016*, 716–722. DOI: 10.1007/978-3-319-30671-1_58. Author PDF.
- [14] Kalervo Järvelin and Jaana Kekäläinen. 2002. “Cumulated Gain-Based Evaluation of IR Techniques.” *ACM Transactions on Information Systems* 20(4):422–446. DOI: 10.1145/582415.582418.
- [15] Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. “SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models.” *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202:38087–38099. PMLR. arXiv: 2211.10438.

Repository evidence cited

- [R1] CHELATEDAI. `chelation_adapter.py`, `BlockAttnResAdapter`, `LayerAttentionAggregator`, and adapter factory. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R2] CHELATEDAI. `config.py`, AttnRes defaults and shallow/balanced/deep presets. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R3] CHELATEDAI. `run_road_course_campaign.py`, AttnRes comparison, depth, trained-depth, and balanced-candidate grids. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R4] CHELATEDAI. `attnres_repeat_seed_decision.py`, aggregate fail-closed decision logic. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R5] CHELATEDAI. “AttnRes Road-Course Benchmark.” 4 May 2026. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R6] CHELATEDAI. “AttnRes Trained Road-Course Benchmark.” 5 May 2026. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R7] CHELATEDAI. “AttnRes NFCorpus Transfer Benchmark.” 5 May 2026. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R8] CHELATEDAI. “AttnRes Repeat-Seed Benchmark.” 5 May 2026. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.
- [R9] CHELATEDAI. `attnres-repeat-seed-decision-2026-05-05.json`, six-run decision summary. Public snapshot `65c9085cd048e8a7351a53e87666fdd5639e612b`. Source.