BIRC: BRIDGING INFORMATION RETRIEVAL AND COMPRESSION THROUGH LATENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent-memory retrieval-augmented generation (RAG) compresses retrieved passages into continuous states, but selecting evidence within a fixed candidate pool restricts access to the rest of the corpus. We introduce BIRC (Bridging Information Retrieval and Compression), a fixed-depth interface in which compressed evidence guides a second corpus search. BIRC uses retrieval relevance to gate passage memories, projects their pooled representation into a feedback query, and orders the final evidence using relevance and source-reconstruction fidelity. Training passes answer-loss gradients through selected-document scores and reconstruction fidelity while keeping passage selection and forward ordering discrete. With IGFR and CFRS settings selected on the evaluation benchmarks, BIRC with Mistral-7B achieves 44.50% average F1 across four open-domain QA benchmarks at $16\times$ compression, exceeding CLaRa by 5.55 percentage points in a complete-system comparison. When both systems are independently trained and evaluated on identical five-passage inputs, BIRC with gating and fidelity-based ordering achieves a 2.90-point advantage with feedback retrieval disabled. BIRC also achieves higher average F1 than CLaRa across all four reader configurations at $16\times$. These gains incur a feedback search and source-reconstruction passes, making BIRC an accuracy–compute trade-off.

1 INTRODUCTION

Retrieval-augmented generation (RAG) gives a reader access to external evidence, at a computational cost that grows with the amount of retrieved text (Lewis et al., 2020; Gao et al., 2023b; Liu et al., 2024; Yoran et al., 2024). Latent-memory methods reduce the reader's input length by encoding each passage as a short sequence of continuous states (Ge et al., 2024; Cheng et al., 2024; He et al., 2025). When these states are used to select evidence from a fixed candidate pool, passages outside that pool remain unavailable as reader input. CLaRa, our primary latent-compression baseline, follows this design: it compresses candidates at a common nominal ratio and uses their representations for passage selection.

BIRC (Bridging Information Retrieval and Compression) connects retrieval and compression through one feedback search per query in its primary configuration (Figure 1). The interface comprises three operations: retrieval scores gate latent memories, pooled gated memories form a feedback query, and source-reconstruction fidelity informs the final evidence order. Adaptive Compression Reweighting (ACR), the Memory-Token Feedback Retrieval Loop (MTFRL), and the Compression-Fidelity Reranking Score (CFRS) implement these operations. Passage selection and evidence ordering remain discrete in the forward pass. MTFRL adds a two-layer feedback projection, the interface's only additional trainable module. The additional search and teacher-forced reconstruction incur online cost: BIRC targets useful evidence under a compressed reader budget, without assuming lower latency than single-pass latent methods.

At $16\times$ with Mistral-7B, BIRC's ACR+CFRS stack exceeds CLaRa by 2.90 points in Avg F1 under independent training on identical five-passage inputs, with supporting retrieval and MTFRL disabled. The native-pipeline difference is 5.55 points. Our contributions are the fixed-depth interface, its selected-document gradient procedure, and controls separating complete-system performance from candidate-source, interface, and memory-allocation effects.

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.

2 Related Work

Adaptive retrieval. Retrieval-augmented generation combines a parametric reader with external evidence (Lewis et al., 2020). Supervised and unsupervised dense retrieval, Fusion-in-Decoder, and joint retriever–reader training provide complementary ways to retrieve and consume that evidence (Karpukhin et al., 2020; Izacard et al., 2022; Izacard & Grave, 2021; Izacard et al., 2023). Adaptive systems vary retrieval timing, route questions by complexity, or alternate between retrieval and reasoning (Jiang et al., 2023c; Jeong et al., 2024; Trivedi et al., 2023; Asai et al., 2024). Classical pseudo-relevance feedback (PRF) revises a query from retrieved documents; HyDE embeds generated hypothetical text (Abdul-Jaleel et al., 2004; Gao et al., 2023a). MemoRAG builds compressed global memory of a long context and decodes draft-answer clues for retrieval (Qian et al., 2025). LycheeMemory recalls compressed KV-cache blocks as its text-based working memory evolves (Chen et al., 2026). BIRC instead builds a query-local memory from five first-pass passages and projects it directly into a corpus-search embedding, without intermediate text generation. Appendix A.24 compares latent feedback with PRF and HyDE under a common 200-candidate follow-up budget; it does not evaluate MemoRAG's global-memory pipeline.

Compression for RAG. Prompt compressors remove or rewrite input tokens (Jiang et al., 2023b; Pan et al., 2024; Jiang et al., 2024), while latent compressors encode a long passage into a small set of continuous states (Chevalier et al., 2023; Mu et al., 2023; Ge et al., 2024; Rau et al., 2025). xRAG maps a frozen retrieval embedding into a single reader token (Cheng et al., 2024); BIRC maps compressed reader-space states back into a retrieval query. PISCO jointly distills its compressor and decoder (Louis et al., 2025). OSCAR and SeleCom make soft compression query-aware (Louis et al., 2026; Liu et al., 2026). SARA combines retained text with compression vectors and uses those vectors for iterative evidence selection within an initially retrieved context set (Jin et al., 2026). CLaRa learns a selector over continuous compressed representations (He et al., 2025), compresses each candidate at a common rate, and selects within the initial candidate pool. BIRC connects these stages through feedback: first-pass scores condition latent memory, the resulting memory drives one additional corpus search, and reconstruction fidelity informs the final passage order.

3 Method: A Fixed-Depth Retrieval–Compression Loop

For query q and corpus $\mathcal{C}$, let $\mathcal{R}(q, \mathcal{C}) = (D_1, \boldsymbol{s}_1)$ denote the supporting retriever's output: five first-pass passages D_1 and their scalar relevance scores $\boldsymbol{s}_1$. Our implementation uses a frozen BGE embedding space and corpus index for feedback retrieval, deterministic MADS–CCEF scoring of the combined candidate set, and a differentiable semantic-score path during Phase II. Appendix A.1 defines these components.

Within the retrieval–compression loop, we truncate each passage to 768 passage-content tokens. Let L_i denote the number of non-structural source tokens retained after truncation. At nominal compression ratio r, the compressor produces $K_i^{(0)} = \max\{1, \lfloor L_i/r \rfloor\}$ ordered states $\mathcal{F}_r(d_i) = (\boldsymbol{m}_i^{(t)})_{t=1}^{K_i^{(0)}}$. BIRC applies the following fixed-depth transition

$$\begin{aligned} (D_1, \boldsymbol{s}_1) &\xrightarrow{\text{score}\to\text{memory}} \widetilde{\boldsymbol{M}}_1 \xrightarrow{\text{memory}\to\text{retrieval}} (D, \boldsymbol{s}), \\ (D, \boldsymbol{s}) &\xrightarrow{\text{score}\to\text{memory}} \widetilde{\boldsymbol{M}} \xrightarrow{\text{fidelity}\to\text{order}} S_{\mathrm{BIRC}}. \end{aligned} \tag{1}$$

Here $\widetilde{\boldsymbol{M}}_1$ and $\widetilde{\boldsymbol{M}}$ collect the gated passage-memory sequences before and after feedback, respectively. BIRC merges and rescores the candidate sets, selects five passages, and reapplies the gate. The score-to-memory and fidelity-to-order operations add no new trainable modules beyond the existing adapters; memory-to-retrieval uses the feedback projection P_{fb}.

Let $\widetilde{\mathcal{F}}_r^{\mathrm{BIRC}}(S)$ denote the reader input formed by recompressing the passages in S, applying the final gates, and concatenating their memories in the order encoded by S. The reader returns

$$\widehat{a} = \mathrm{GreedyDecode}_\theta\Big(q, \widetilde{\mathcal{F}}_r^{\mathrm{BIRC}}(S_{\mathrm{BIRC}}(q, \mathcal{C}))\Big). \tag{2}$$

At evaluation, greedy decoding selects the highest-probability token at each step and stops at the end-of-sequence token or after 64 newly generated tokens. Figure 1 summarizes this interface-level view, and Appendix A.1 specifies the supporting retriever used in the native-pipeline experiments.

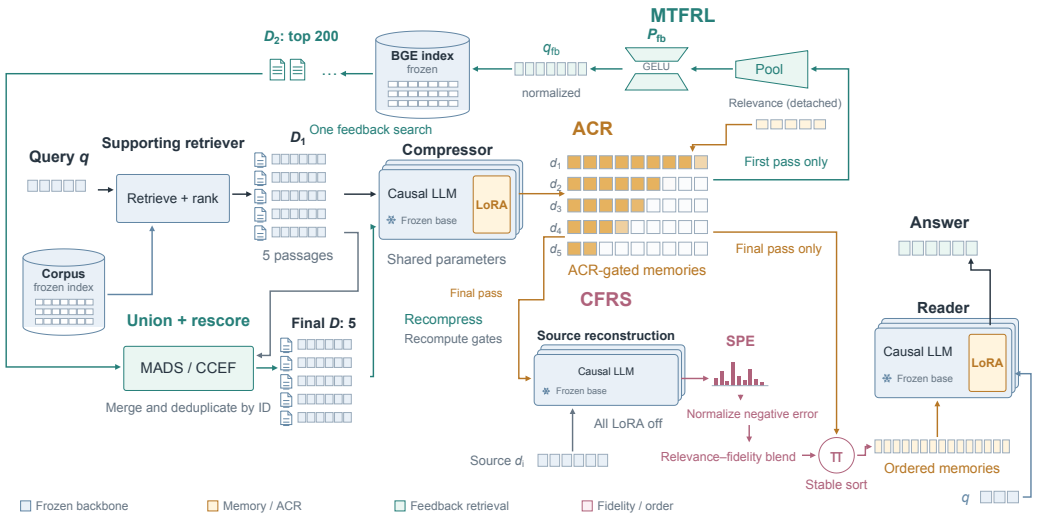

Figure 1: BIRC overview: relevance gates memory (ACR), pooled memory drives one feedback retrieval (MTFRL), and reconstruction fidelity informs final ordering (CFRS). Memory shading and probability bars are schematic.

3.1 Score-to-memory: Adaptive Compression Reweighting

ACR maps the fused scores for the current five passages to gate scales $\rho_i \in [\rho_{\min}, \rho_{\max}] = [0.25, 1.0]$:

$$\rho_i = \rho_{\min} + (\rho_{\max} - \rho_{\min}) \frac{s_i - \min_j s_j}{\max_j s_j - \min_j s_j + \epsilon}, \tag{3}$$

where $\epsilon = 10^{-6}$. Each scale defines a soft cutoff along the ordered memory sequence,

$$g_i^{(t)} = \sigma\Big(\beta(\rho_i K_i^{(0)} - t)\Big), \qquad \widetilde{\boldsymbol{m}}_i^{(t)} = g_i^{(t)} \boldsymbol{m}_i^{(t)}, \tag{4}$$

where σ is the logistic sigmoid, $\beta = 10$, and $t \in \{1, \ldots, K_i^{(0)}\}$. The scores used to compute these scales are detached from the gradient computation. Equal scores produce equal scales; passages with the minimum score receive $\rho_{\min}$, including all five passages in the equal-score case. We recompute the scales after feedback retrieval updates the five-passage set.

Gating scales each state without changing the memory sequence length. We define

$$T_i = \max\Big\{1, \Big\lfloor \rho_i K_i^{(0)} \Big\rfloor\Big\} \tag{5}$$

as the number of first-pass memory positions pooled by MTFRL. Recomputed T_i values on the final set summarize the gate profile; the reader still receives all base memory positions. Physical sequence length, KV-cache use, and latency are reported in Appendix A.42. ACR is an input-state gate: RMSNorm can cancel positive scalar scaling before the first attention projections. Its remaining effects act through residual states, normalization's numerical floor, and feedback pooling, as derived in Appendix A.49.1. It does not guarantee attention suppression.

3.2 Memory-to-retrieval: Memory-Token Feedback Retrieval

MTFRL pools the gated first-pass memories and projects them into the normalized BGE embedding space:

$$\overline{\boldsymbol{m}} = \mathrm{Normalize}\left(\frac{1}{5}\sum_{i=1}^{5}\frac{1}{T_i}\sum_{t=1}^{T_i}\widetilde{\boldsymbol{m}}_i^{(t)}\right), \qquad \boldsymbol{q}_{\mathrm{fb}} = \mathrm{Normalize}(P_{\mathrm{fb}}(\overline{\boldsymbol{m}})). \tag{6}$$

Here $\mathrm{Normalize}$ denotes ℓ_2 normalization, and P_{fb} is a two-layer GELU projection into the 1,024-dimensional BGE space, initialized from a pre-fitted query-space map (Appendix A.2). The initialization supplies a coordinate map, while Phase II adapts it to pooled memories. The feedback

query retrieves 200 candidates; BIRC deduplicates and rescores their union with D_1 and retains five passages. The outer mean gives each passage equal coefficient, not relevance-proportional weight; noisy first-pass passages can therefore distort feedback (Appendix A.49.2). Primary results use one feedback retrieval; Appendix A.25 varies test-time depth with parameters fixed.

3.3 Fidelity-to-order: Compression-Fidelity Reranking

CFRS uses source reconstructability as a proxy for retained information, complementing query relevance; predictable text need not support the answer. For each of the final five passages, it adds a teacher-forced source pass through the frozen decoder with all LoRA adapters disabled. This cost occurs at inference as well as training. During Phase II, gradients through the ACR-gated memories still reach the compressor adapter. Let $e_i = \mathrm{SPE}(d_i)$ denote the squared probability error: the squared ℓ_2 distance between the next-token probability vector and the one-hot source target, averaged over retained source-token positions. SPE measures prediction error in probability space; the hidden-state alignment loss below operates on reader representations. CFRS negates this error and min–max normalizes the result:

$$f_i = \frac{-e_i - \min_{d_j \in D}(-e_j)}{\max_{d_j \in D}(-e_j) - \min_{d_j \in D}(-e_j) + 10^{-6}}. \tag{7}$$

Both $s_{\mathrm{fused}}(d_i)$ and f_i lie in $[0, 1]$. CFRS combines them as

$$s_{\mathrm{final}}(d_i) = 0.7\, s_{\mathrm{fused}}(d_i) + 0.3\, f_i. \tag{8}$$

A stable descending sort gives the inference-time order. Training retains this hard order in the forward pass while a row-wise soft assignment carries QA gradients through fidelity. With $u_i = s_{\mathrm{final}}(d_i)$ and π_r the document at hard rank r,

$$P^{\mathrm{soft}}_{ri} = \mathrm{softmax}_i\left(-\frac{(u_i - \mathrm{sg}(u_{\pi_r}))^2}{\tau_{\mathrm{ord}}}\right), \quad P = P^{\mathrm{hard}} + P^{\mathrm{soft}} - \mathrm{sg}(P^{\mathrm{soft}}), \tag{9}$$

where $\tau_{\mathrm{ord}} = 0.1$ is the ordering temperature and sg denotes the stop-gradient operator. Let $\widetilde{\boldsymbol{M}}$ stack the five padded passage-memory tensors; $P\widetilde{\boldsymbol{M}}$ permutes their document dimension. The normalization minimum and span are detached, so gradients reach the compressor through $s_{\mathrm{final}} \to f_i \to e_i$ without changing the hard forward order. There is no stop-gradient for demoted passages. Appendix A.49.3 derives the sorting gradient and its possible conflict with reconstruction.

3.4 Training

Phase I trains the compressor adapter to predict paired target sequences from memory alone:

$$\mathcal{L}_{\mathrm{target}} = -\mathbb{E}_{(d,y)\sim\mathcal{D}_{\mathrm{pre}}} \log p_\theta(y \mid \mathcal{F}_r(d)). \tag{10}$$

The compressor receives d, and the decoder predicts y from its memory. The fixed Phase-I mixture contains 7,808,465 pairs across four functional categories: SimpleQA, ComplexQA, Paraphrase, and Entity-Augmented (Appendix A.17). Phase II minimizes

$$\mathcal{L} = \mathcal{L}_{\mathrm{QA}} + 0.1\, \mathcal{L}_{\mathrm{HSA}}, \tag{11}$$

with

$$\mathcal{L}_{\mathrm{QA}} = -\mathbb{E} \log p_\theta\Big(a^* \mid q, \widetilde{\mathcal{F}}^{\mathrm{BIRC}}_r(S_{\mathrm{BIRC}}(q, \mathcal{C}))\Big), \tag{12}$$

and

$$\mathcal{L}_{\mathrm{HSA}} = \mathbb{E}\left[\|\overline{\boldsymbol{h}}_{\mathcal{M}} - \overline{\boldsymbol{h}}_{\mathcal{P}_{\mathrm{QA}}}\|_2^2\right]. \tag{13}$$

Here $\mathcal{M}$ denotes memory positions, and $\mathcal{P}_{\mathrm{QA}}$ denotes the union of non-structural query and gold-answer token positions in the same QA forward pass. Both means use the reader's final hidden layer and both remain in the computation graph. This symmetric regularizer sums over hidden dimensions and averages over examples; the query–answer mean is not a fixed teacher. QA cross-entropy constrains prediction, but does not prove non-collapse. The independently retrained Stop-QA control achieves 44.55 Avg F1 versus 44.50 for the primary objective (Appendix A.12).

Phase II retains the trained compressor adapter and copies its weights to initialize the query-reasoner adapter. It jointly trains rank-16 LoRA adapters (Hu et al., 2022) for query reasoning, compression,

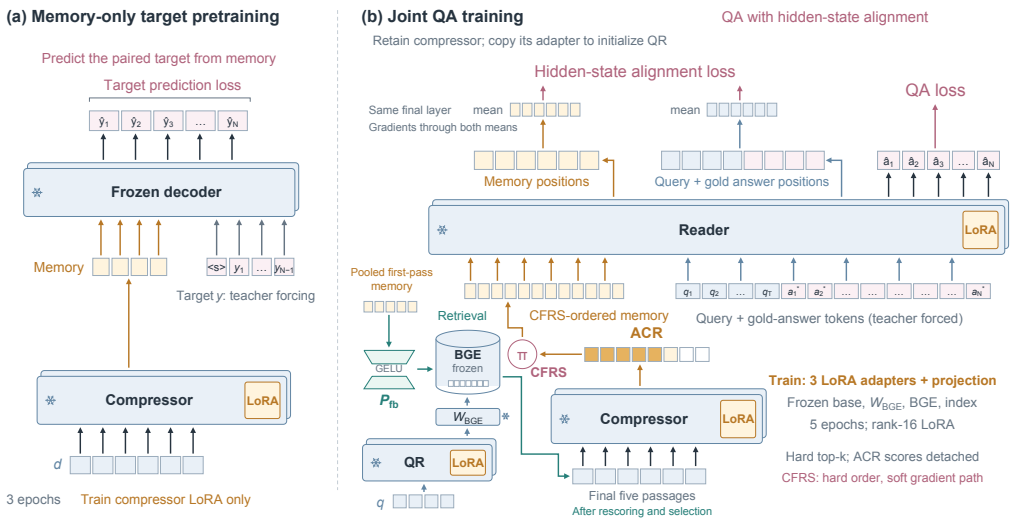

Figure 2: Two-phase training: memory-only target pretraining, followed by joint QA and hidden-state alignment training of three LoRA adapters and the feedback projection. The base model, BGE retriever, and index remain frozen.

and reading, together with the feedback projection P_{fb}. The base-model weights, BGE retriever, and corpus index remain frozen. Phase I runs for three epochs and Phase II for five epochs. A separate model is trained for each $r \in \{4, 16, 32, 64, 128\}$.

Training uses straight-through score surrogates for selected documents while keeping top-k membership discrete. Let s^{stored} and s^{live} denote the stored fused score and its live MADS–CCEF value, whose differentiable component is the BGE cosine (Appendix A.1). The surrogate is $s^{\text{stored}} + s^{\text{live}} - \text{sg}(s^{\text{live}})$. For a document introduced by feedback retrieval, an analogous term $z - \text{sg}(z)$, where z is its feedback cosine, supplies a gradient to P_{fb} without changing the forward score. Appendix A.1 gives the full algorithm.

4 Experiments

4.1 Datasets and metrics

Mistral-7B-Instruct-v0.2 is the primary reader (Jiang et al., 2023a); cross-reader experiments use Llama-3-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-14B-Instruct (Qwen et al., 2024). The benchmarks are NQ (Kwiatkowski et al., 2019), HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and 2WikiMultiHopQA (Ho et al., 2020). Our primary evaluator computes token-level F1 after normalization and whitespace tokenization against the single reference answer stored in each prepared benchmark record. We use the same evaluator and reference answers for BIRC, CLaRa, all ablations, and the external predictions rescored in this work. Rows taken from the literature retain their source's evaluation convention, as specified in the corresponding table captions. Appendix A.44 defines Contains Exact Match (CEM), Exact Match (EM), and supporting-page Recall@k; Appendix A.45 evaluates Full BIRC with alias-aware scoring. "Avg" denotes the unweighted mean of the four benchmark scores.

4.2 Comparison design

The Normal retrieval setting searches KILT Wikipedia (Petroni et al., 2021). BIRC's supporting retriever forms a fused top-4,000 sparse–dense candidate pool, applies optional multi-hop coverage repair and fixed score fusion, and selects five first-pass passages; the interface then performs one 200-candidate feedback search. CLaRa selects five passages from a frozen-BGE top-20 pool. Oracle retrieval supplies both systems with the same pool of passages from 100 distinct pages, including annotated support pages without a rank bonus. Appendix A.1 describes the pool construction.

The native comparison matches reader, corpus, Phase II examples, schedule, nominal ratio, and seeds while retaining each method's retrieval procedure, objective, adapters, and initialization. It is therefore a complete-system comparison. The controlled experiments use shared candidates, matched selector input sizes, component swaps, matched physical memory lengths, a 2×2 factorial design, and text-space feedback substitutions. Each experiment specifies its training and evaluation conditions.

Selection and evaluation. The IGFR iteration caps and CFRS weight were selected using the benchmark evaluation results, including $K_{\text{igfr}} = 1$ at $128\times$ and $w_{\text{cfrs}} = 0.30$. The separate development search covers only the CCEF agreement floor and ACR parameters (Appendix A.15). Reported scores and intervals are therefore conditional on evaluation-set selection, including controls that reuse these settings; they are not estimates from an untouched test set. Bootstrap tests and seed intervals do not correct this selection bias.

Phase II uses $38{,}400$ benchmark-balanced examples per epoch for five epochs. Unless noted otherwise, results average seeds $\{42, 123, 456, 789, 2024\}$. Seed-level intervals use Student's t with four degrees of freedom; query-level paired comparisons use a 10,000-resample bootstrap (Berg-Kirkpatrick et al., 2012). Tables identify the interval type and Holm-adjustment family. Appendix A.47 details aggregation and resampling. Statistics precede display rounding.

5 Results

5.1 Controlled comparisons of the three interface updates

Table 1 compares the three interface operations using fixed-checkpoint interventions and independent retraining. With fixed checkpoints, Full BIRC and the configuration with all three updates disabled achieve Avg F1 scores of 44.50 and 41.30, respectively. After independent retraining, Full BIRC exceeds the variants without fidelity-to-order, with uniform gating, and with original-query feedback by 0.55, 0.72, and 0.67 percentage points in Avg F1, respectively. Its margin over the all-updates-off configuration is 2.95 points.

Table 1: Interface controls ($16\times$, Normal; five-seed means, %). Fixed-checkpoint changes apply only at evaluation; other controls are retrained. MTFRL control: no feedback versus original-query feedback, respectively. All-off: bypass versus uniform gates, with CFRS removed. Paired Avg F1 CIs use Student's t (4 df); bold marks maxima. Protocols: Appendices A.19 and A.20.

	Fixed checkpoint		**Independent retraining**		**Retrained Full−variant**
Configuration	**Avg CEM**	**Avg F1**	**Avg CEM**	**Avg F1**	**95% CI**
Full BIRC	**41.30**	**44.50**	**41.20**	**44.47**	—
CFRS-free	40.55	43.70	40.65	43.92	$[+0.47, +0.63]$
Uniform gate	40.35	43.55	40.48	43.75	$[+0.58, +0.86]$
MTFRL control	40.42	43.62	40.54	43.80	$[+0.56, +0.78]$
All updates off	39.10	41.30	39.30	41.52	$[+2.76, +3.14]$

Appendix Table A24 extends the fixed-checkpoint controls across all five compression ratios, and Appendix A.20 reports the complete independently retrained configurations.

5.2 Controlled comparisons across the retrieval–compression boundary

Table 2(a) crosses retrieval and compressor–reader stacks. BIRC first-pass retrieval with the CLaRa stack reaches 41.24 Avg F1 without updates and 43.30 with CFRS; CLaRa retrieval with the BIRC ACR+CFRS stack reaches 41.55. Both swaps disable MTFRL and exchange only passage identities, text, and scores through their respective five-passage selectors.

Panel (b) independently trains both systems on the same frozen-BGE top-five passages, disabling BIRC's supporting retriever and feedback search. BIRC and CLaRa achieve Avg F1 scores of 41.35 and 38.45, respectively. Across five paired training seeds, the mean difference is 2.90 points (sample SD 0.29), with a paired 95% CI of $[+2.54, +3.26]$ (Student's t, 4 df; per-seed values in Appendix A.23).

Table 2: Two controls (16×, Normal; five-seed means, %). (a) Retrieval/stack swaps; only Full BIRC uses MTFRL. (b) Shared frozen-BGE top-five passages throughout Phase II and evaluation, with BIRC retrieval and feedback disabled. Δ: BIRC minus CLaRa (pp). Bold marks maxima and Δ. Protocols: Appendices A.21 and A.23.

(a) Conditional component swaps

Configuration	Retrieval	Comp.–reader	Updates	Avg F1
CLaRa	CLaRa	CLaRa	*N/A (native)*	38.95
CLaRa-ret → BIRC-stack	CLaRa	BIRC	ACR+CFRS	41.55
BIRC-ret → CLaRa-stack	BIRC	CLaRa	Off	41.24
BIRC-ret → CLaRa-stack	BIRC	CLaRa	CFRS	43.30
Full BIRC	BIRC	BIRC	Full	**44.50**

(b) Shared fixed top-five candidates from the frozen BGE retriever

System	Avg EM	Avg CEM	Avg F1	Δ Avg F1
CLaRa	29.83	38.16	38.45	—
BIRC stack (ACR+CFRS; MTFRL off)	**32.33**	**39.85**	**41.35**	**+2.90**

5.3 Native-pipeline performance across compression ratios

Across compression ratios. Table 3 reports Normal-retrieval results at five compression ratios; Appendix A.7 gives the complete Normal- and Oracle-retrieval results. BIRC's Avg F1, averaged over five seeds, is higher than CLaRa's at every reported operating point; 128× uses a different IGFR cap (Appendix A.28). At 16×, the seed-level 95% CIs for Avg F1 are $[44.21, 44.79]$ for BIRC and $[38.69, 39.21]$ for CLaRa. The mean paired-seed difference is 5.55 pp, with a 95% CI of $[5.39, 5.71]$ (Student's t, 4 df); Appendix A.30 reports all benchmark-, metric-, and seed-level values.

Table 3: Native-pipeline results (Normal; five-seed means, %; evaluation-selected BIRC settings). Retrieval and candidate budgets are system-specific; CR is nominal. Bold marks within-CR maxima. Full results and the 128× IGFR setting: Appendices A.7 and A.28.

CR	System	NQ F1	Hotpot F1	MuSiQue F1	2Wiki F1	Avg F1	Avg CEM
4×	BIRC	**56.18**	**54.23**	**22.37**	**50.94**	**45.93**	**42.15**
	CLaRa	48.34	46.07	17.55	47.32	39.82	39.98
16×	BIRC	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**	**41.30**
	CLaRa	47.94	45.85	17.68	44.33	38.95	38.85
32×	BIRC	**54.87**	**49.63**	**19.27**	**47.11**	**42.72**	**39.15**
	CLaRa	47.28	44.28	16.72	42.82	37.77	37.90
64×	BIRC	**53.06**	**49.87**	**19.13**	**44.82**	**41.72**	**38.18**
	CLaRa	46.41	44.65	16.08	40.57	36.93	37.65
128×	BIRC	**51.83**	**47.96**	**18.07**	38.02	**38.97**	**36.95**
	CLaRa	45.54	42.29	14.50	**38.31**	35.16	36.75

Other native systems. In the five-seed comparison with matched training schedules, each method retains its native retrieval path, compression mechanism, objective, adapter placement, and initialization. BIRC's Avg F1 exceeds SARA, SeleCom, and OSCAR by 1.20, 1.95, and 2.55 points, respectively (all Holm-adjusted $p \leq 0.003$; Appendix A.9). These are system-level differences: shared-candidate retraining against these three methods is not available to isolate compression from retrieval. A separate unified evaluator comparison retains each method's native training and seed count (Appendix A.8). Alias-aware scoring of fixed BIRC outputs raises Avg EM/CEM/F1 by $0.88/0.78/0.43$ points (Appendix A.45).

Across readers. Across Mistral-7B, Llama-3-8B, and Qwen2.5-14B at 16×, BIRC has higher F1 than CLaRa in all twelve reader–benchmark combinations (Table 4). Each pair shares reader, training

schedule, and seeds while retaining native retrieval. Appendices A.18 and A.34 give complete results. A separate evaluation with a frozen, 4-bit-quantized Llama-3-70B-Instruct reader capped at 1,024 input tokens yields Avg F1 scores of 51.40% for BIRC and 46.80% for CLaRa (Appendix A.35).

Table 4: Native-pipeline F1% across readers ($16\times$, Normal; five-seed means, BIRC/CLaRa), with system-specific retrieval and budgets. Δ Avg F1: BIRC minus CLaRa (pp). Bold marks pairwise maxima.

Reader backbone	NQ	HotpotQA	MuSiQue	2Wiki	Avg	Δ Avg F1
Mistral-7B	**55.19**/47.94	**52.71**/45.85	**21.30**/17.68	**48.80**/44.33	**44.50**/38.95	+5.55
Llama-3-8B	**58.63**/50.37	**55.76**/48.41	**24.07**/20.18	**51.94**/46.80	**47.60**/41.44	+6.16
Qwen2.5-14B	**55.87**/50.15	**53.36**/47.99	**22.68**/19.76	**50.49**/46.54	**45.60**/41.11	+4.49

5.4 Memory allocation controls

At $16\times$, the mean gate scale/prefix count falls from $0.98/36.1$ at rank 1 to $0.31/11.4$ at rank 5 (Table 5(a)). The reader receives all 184 base positions on average; the mean 108.0 prefix total is diagnostic. Panel (b) instead physically allocates an identical number of positions per query under ACR or a uniform rule, on the same first-pass passages with MTFRL disabled. Each pair uses one physical-ACR-trained checkpoint. Physical ACR gains 0.80 Avg F1 at essentially equal measured latency and memory.

Table 5: Gate diagnostics and physical control (nominal $16\times$). (a) 1,000-query statistics; L/T_i uses rounded $L = 590$, with all input states materialized. (b) Length-matched allocation on the same five first-pass passages, without feedback. Task metrics: %; Δ: ACR minus uniform.

(a) Score-conditioned ACR gate statistics

Final rank	Mean ρ_i	Mean T_i	Diagnostic L/T_i (accounting only)
1	0.98	36.1	$16.3\times$
2	0.74	27.2	$21.7\times$
3	0.52	19.1	$30.9\times$
4	0.38	14.0	$42.1\times$
5	0.31	11.4	$51.8\times$

(b) Matched physical-position control

Allocation	Avg EM	Avg CEM	Avg F1	Mean positions	Sec./query	Peak GB
Physical ACR	**35.40**	**41.15**	**44.40**	108.0	6.15	20.6
Matched uniform	34.70	40.45	43.60	108.0	6.14	20.6
Δ (ACR$-$uniform)	**+0.70**	**+0.70**	**+0.80**	0.0	+0.01	0.0

5.5 Answer quality and computational cost

BIRC's batch-1 native profile on one A100-80GB takes 6.62 s/query and peaks at 21.2 GB. Its compression/CFRS and reader stages total $2.68 + 3.32 = 6.00$ s; CLaRa's selection/compression and reader stages total $2.54 + 3.29 = 5.83$ s. This 0.17 s stage-subtotal difference excludes BIRC's 0.62 s routing/scoring cost. It is not a matched end-to-end rerun: the candidate paths and query-encoding attribution differ (Appendix A.42).

First-pass direct context achieves 45.40 Avg F1, versus BIRC's 44.50 (Appendix A.26), but its latency and peak memory are not reported. The 8.43 s/32.4 GB frozen-BGE direct-context profile uses different passages and cannot establish BIRC's efficiency relative to that stronger endpoint. Matched within-BIRC controls quantify feedback and physical-allocation costs (Appendices A.41 and A.42).

6 DISCUSSION

The retrained single-update drops sum to 1.94 points, below the 2.95-point all-updates-off drop. Gating, feedback retrieval, and ordering may therefore complement one another, but the available configurations do not identify a three-way interaction. Candidate-source and stack contrasts remain conditional on their training protocols and evaluation-selected settings.

High compression also increases sensitivity to gating. At $128\times$, raising the gate floor to 0.45 for multi-hop and multi-aspect queries changes Avg CEM from 36.95 to 37.60 in a fixed-checkpoint sensitivity study (Appendix A.33). This exploratory intervention is separate from the primary common-floor configuration.

7 CONCLUSION

BIRC links score-gated memory, one feedback search, and fidelity-based ordering. At $16\times$, its Mistral-7B stack exceeds CLaRa by 2.90 Avg F1 points on shared candidates; the native-pipeline difference is 5.55 points. These describe configurations with evaluation-selected IGFR caps and CFRS weight. Independent evaluation with development-selected settings is needed to establish the magnitude of the generalization gain.

LIMITATIONS

BIRC uses a fixed feedback depth and frozen BGE index. ACR retains all physical positions, while local score normalization and unweighted feedback pooling can amplify retrieval noise. Source fidelity is an indirect ordering signal, and neither HSA nor CFRS supplies a guarantee of information preservation. Our English QA results do not establish exact retention of rare names or numbers, or robustness to counterfactual evidence. Matched accuracy–latency evaluations against xRAG and MemoRAG, and shared-candidate controls for SARA, SeleCom, and OSCAR, remain outside the evidence here. Appendix A.48 also specifies the hyperparameter-provenance and training-overlap audit limits.

AI USE STATEMENT

Generative AI tools assisted with manuscript wording and code implementation. The authors reviewed all AI-assisted material and take responsibility for the submitted manuscript and code.

ETHICS STATEMENT

The study uses released QA benchmarks and public KILT Wikipedia text and does not collect new sensitive participant data. We used canonical evaluation-page URLs and IDs to exclude exact page matches from both Phase-I and Phase-II training; Appendix A.46 reports the expanded overlap audit and Dedup+ sensitivity. The derived MuSiQue training partition contains no page URLs matching those in the test splits and remains subject to the source dataset's CC BY 4.0 license. Appendix A.47 lists the dataset and corpus licenses.

REFERENCES

Nasreen Abdul-Jaleel, James Allan, W. Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Donald Metzler, Mark D. Smucker, Trevor Strohman, Howard Turtle, and Courtney Wade. UMass at TREC 2004: Novelty and HARD. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*. National Institute of Standards and Technology, 2004. URL https://trec.nist.gov/pubs/trec13/papers/umass.novelty.hard.pdf.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. An empirical investigation of statistical significance in NLP. In Jun'ichi Tsujii, James Henderson, and Marius Paşca (eds.), *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 995–1005, Jeju Island, Korea, July 2012. Association for Computational Linguistics. URL https://aclanthology.org/D12-1091/.

Zhuoen Chen, Dongfang Li, Meishan Zhang, Baotian Hu, and Min Zhang. Dynamic long context reasoning over compressed memory via end-to-end reinforcement learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8064–8083, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.365. URL https://aclanthology.org/2026.acl-long.365/.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. xRAG: Extreme context compression for retrieval-augmented generation with one token. In *Advances in Neural Information Processing Systems*, volume 37, pp. 109487–109516. Curran Associates, Inc., 2024. doi: 10.52202/079017-3476.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. Adapting language models to compress contexts. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3829–3846, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.232. URL https://aclanthology.org/2023.emnlp-main.232/.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.99. URL https://aclanthology.org/2023.acl-long.99/.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2023b. URL https://arxiv.org/abs/2312.10997.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. In-context autoencoder for context compression in a large language model. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=uREj4ZuGJE.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jie He, Richard He Bai, Sinead Williamson, Jeff Z. Pan, Navdeep Jaitly, and Yizhe Zhang. CLaRa: Bridging retrieval and generation with continuous latent reasoning, 2025. URL https://arxiv.org/abs/2511.18659v3.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580/.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python, 2020. URL https://doi.org/10.5281/zenodo.1212303.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.74. URL https://aclanthology.org/2021.eacl-main.74/.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=jKN1pXi7b0.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023. URL https://jmlr.org/papers/v24/23-0037.html.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023a. URL https://arxiv.org/abs/2310.06825.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.825. URL https://aclanthology.org/2023.emnlp-main.825/.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.91. URL https://aclanthology.org/2024.acl-long.91/.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495/.

Yiqiao Jin, Kartik Sharma, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. SARA: Selective and adaptive retrieval-augmented generation with context compression. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14508–14528, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.661. URL https://aclanthology.org/2026.acl-long.661/.

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 Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

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/.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Yunhao Liu, Zian Jia, Xinyu Gao, Kanjun Xu, and Yun Xiong. Rethinking soft compression in retrieval-augmented generation: A query-conditioned selector perspective. In *Proceedings of the ACM Web Conference 2026*, pp. 2441–2452. ACM, April 2026. doi: 10.1145/3774904.3792700. URL https://doi.org/10.1145/3774904.3792700.

Maxime Louis, Hervé Déjean, and Stéphane Clinchant. PISCO: Pretty simple compression for retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 15506–15521, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.800. URL https://aclanthology.org/2025.findings-acl.800/.

Maxime Louis, Thibault Formal, Hervé Déjean, and Stéphane Clinchant. OSCAR: Online soft compression for RAG. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 3102–3121, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/064aae662bdd38d031e33596b215d50c-Paper-Conference.pdf.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Jesse Mu, Xiang Lisa Li, and Noah Goodman. Learning to compress prompts with gist tokens. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 19327–19352. Curran Associates, Inc., 2023. doi: 10.52202/075280-0848. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3d77c6dcc7f143aa2154e7f4d5e22d68-Paper-Conference.pdf.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Dongmei Zhang. LLMLingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.57. URL https://aclanthology.org/2024.findings-acl.57/.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2523–2544, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.200. URL https://aclanthology.org/2021.naacl-main.200/.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. MemoRAG: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pp. 2366–2377. ACM, April 2025. doi: 10.1145/3696410.3714805. URL https://doi.org/10.1145/3696410.3714805.

Qwen, An Yang, Baosong Yang, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. doi: 10.1162/tacl_a_00605. URL https://aclanthology.org/2023.tacl-1.75/.

David Rau, Shuai Wang, Hervé Déjean, Stéphane Clinchant, and Jaap Kamps. Context embeddings for efficient answer generation in retrieval-augmented generation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pp. 493–502. ACM, March 2025. doi: 10.1145/3701551.3703527. URL https://doi.org/10.1145/3701551.3703527.

Zhengliang Shi, Shuo Zhang, Weiwei Sun, Shen Gao, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. Generate-then-ground in retrieval-augmented generation for multi-hop question answering. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7339–7353, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.397. URL https://aclanthology.org/2024.acl-long.397/.

Zhengliang Shi, Lingyong Yan, Weiwei Sun, Yue Feng, Pengjie Ren, Xinyu Ma, Shuaiqiang Wang, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. Direct retrieval-augmented optimization: Synergizing knowledge selection and language models. *ACM Transactions on Information Systems*, 44(4):1–30, May 2026. doi: 10.1145/3795527. URL https://doi.org/10.1145/3795527.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. ♫ MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10014–10037, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.557. URL https://aclanthology.org/2023.acl-long.557/.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649, Washington, DC, USA, July 2024. ACM. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. RECOMP: Improving retrieval-augmented LMs with context compression and selective augmentation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZS4m74kZpH.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1910.07467.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A APPENDIX

This appendix describes the supporting retriever and optimization procedure and reports additional QA results, controlled comparisons, sensitivity analyses, reader evaluations, execution profiles, and overlap audits. Unless otherwise stated, experiments use the protocol in §4.

Reading map. Sections A.1, A.2, and A.47 specify the retriever and training configuration. Sections A.7–A.21 report additional results and interface-update controls. Sections A.22 and A.23 report the selector-input-size and shared-candidate controls. Section A.12 reports controls for HSA reduction, gradient routing, and CCEF selection policies. Later sections provide sensitivity analyses, reader evaluations, computational measurements, and overlap audits.

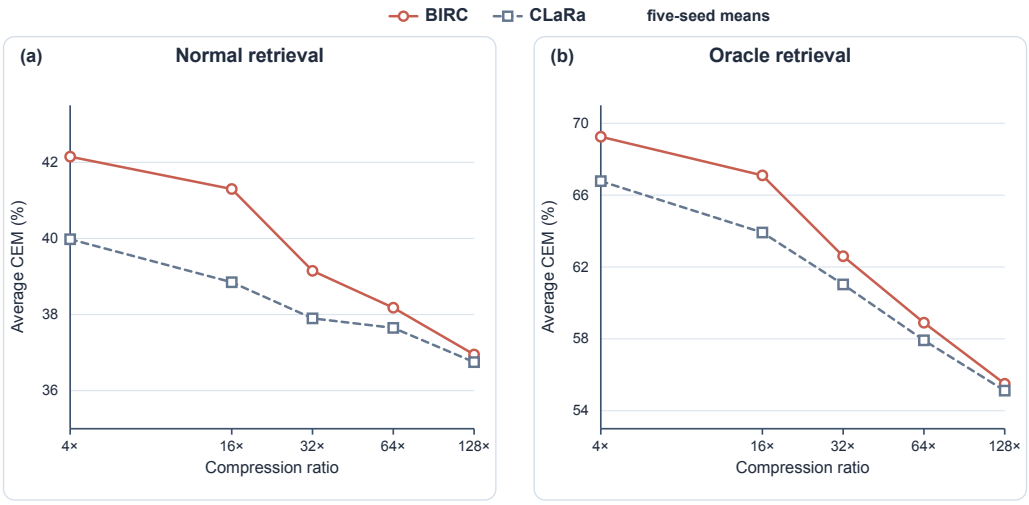

Figure A1: Average CEM (%) across compression ratios under Normal and Oracle retrieval (Mistral-7B; five-seed means).

A.1 SUPPORTING RETRIEVER AND INTERFACE IMPLEMENTATION

The supporting operator $\mathcal{R}$ supplies BIRC with five passage identities, their text, and scalar relevance scores. The native-pipeline experiments instantiate $\mathcal{R}$ with deterministic query routing, sparse–dense acquisition, optional multi-hop coverage repair, and fixed score fusion. These steps are implemented by Question Complexity Assessment (QCA), Adaptive Hybrid Retrieval (AHR), Iterative Gap-Filling Retrieval (IGFR), Multi-Agent Document Scoring (MADS), and Confidence-Calibrated Evidence Fusion (CCEF). MADS uses three deterministic scorers, and CCEF measures their score agreement. BIRC's three directed updates begin after $\mathcal{R}$ returns $(D_1, \boldsymbol{s}_1)$.

For query q, let $D_q(r)$ be the final five-passage set. Let L_{qi} denote the number of non-structural source tokens retained for passage i under the 768-token passage-content limit, and let $T_{qi}^{\text{final}}(r)$ denote the gate-induced prefix statistic recomputed on the final set. These final-set counts are accounting statistics; only the corresponding first-pass counts in Eq. 6 determine feedback pooling. We summarize the gate profile by

$$r_{\text{gate}}(r) = \frac{\sum_q \sum_{i \in D_q(r)} L_{qi}}{\sum_q \sum_{i \in D_q(r)} T_{qi}^{\text{final}}(r)}. \tag{14}$$

CFRS changes order but not set membership.

The query reasoner uses the final-layer hidden state at the last non-padding input position under the backbone's native tokenizer as $\boldsymbol{q}_{\text{rep}}$. The integrated compressor uses final-layer states at the inserted <MEM> positions as its ordered memory sequence.

A.1.1 DETERMINISTIC QUERY ROUTING (QCA)

QCA assigns each query q to Simple, Multi-Aspect, or Multi-Hop using 38 surface-level rules (14 Multi-Hop, 12 Multi-Aspect, and 12 Simple) and named-entity counts from spaCy (Honnibal et al., 2020). Each rule ℓ contributes an indicator $I_\ell(q) \in \{0, 1\}$ and weight w_ℓ, giving $c(q) = \sum_\ell w_\ell I_\ell(q) / \sum_\ell w_\ell$. The denominator is the total weight of the 38-rule inventory, including inactive rules, so $c(q)$ is the fraction of the inventory weight activated by q. This attenuates routing toward balanced sparse–dense weights; it is not a calibrated probability of query complexity. Normalizing only over active rules would give $c(q) = 1$ whenever any rule fires and would define a different operator. AHR uses $c(q)$ to interpolate toward the routed target pair. A query is routed to Multi-Hop when a hop rule fires and the query contains at least two named entities; it is routed to Multi-Aspect when an aspect rule fires and no hop rule fires; otherwise it is Simple. Hop rules take precedence over aspect rules. The routed category selects the type-specific AHR defaults, while $c(q)$ controls the weighted rule-activation interpolation. For Multi-Hop queries, QCA also produces deterministic subquestions for IGFR using syntactic templates, with no LLM generation. Appendix A.3 lists the complete rule inventory and fixed weights. Fixed-checkpoint and zero-shot-LLM routing comparisons are reported in Appendices A.3 and A.38.

A.1.2 SUPPORTING-RETRIEVER IMPLEMENTATION

Sparse–dense candidate acquisition (AHR). AHR combines BM25 scores with dense similarities from BGE-large-en-v1.5 (Xiao et al., 2024). Its dense query uses the fixed W_{BGE} projection. The weight pair $(w_{\mathrm{bm25}}, w_{\mathrm{dense}})$ is type-conditioned with defaults $(0.75, 0.25)$/$(0.30, 0.70)$/$(0.25, 0.75)$ for Simple/Multi-Aspect/Multi-Hop, respectively. The realized weights use continuous interpolation based on the weighted rule-activation score: $\boldsymbol{w}(q) = (1 - c(q))(0.5, 0.5) + c(q)\boldsymbol{w}_{\mathrm{type}}$. For $0 < c(q) < 1$, the realized weights lie strictly between the balanced and type-specific pairs; the interpolation is continuous even though the routed category is discrete. Table A17 reports the corresponding downstream sensitivity. Each retrieval branch returns its top $4{,}000$ results, and its scores are min–max normalized within that set. For a document absent from a nonempty branch, its score in that branch is set to the branch minimum; a constant-score or empty branch contributes 0.5. AHR then mixes the two scores and retains the top $4{,}000$; ties are resolved by corpus order and then document ID.

Optional multi-hop coverage repair (IGFR). For Multi-Hop queries, IGFR computes entity coverage $\mathrm{cov}(q, D) = |\mathcal{E}_q \cap \mathcal{E}_D| / |\mathcal{E}_q|$. If $\mathcal{E}_q$ is empty, coverage is defined as 1, so IGFR is not triggered. When coverage falls below $\gamma = 0.5$, it issues targeted subqueries from QCA's deterministic decomposition. Each round retrieves up to 200 candidates, merges them with the AHR pool, removes duplicates, and stops when coverage reaches 0.5, the template list is exhausted, or the iteration limit is reached. We set $K_{\mathrm{igfr}} = 2$ at $4\times$–$64\times$ and $K_{\mathrm{igfr}} = 1$ at $128\times$. The decomposition uses deterministic surface rules rather than the model-generated chain-of-thought–retrieval interleaving of IRCoT (Trivedi et al., 2023).

Candidate scoring (MADS). Three scoring modules—Lexical $\mathcal{S}_{\mathrm{lex}}$ (TF–IDF cosine), Semantic $\mathcal{S}_{\mathrm{sem}}$ (BGE embedding cosine), and Entity $\mathcal{S}_{\mathrm{ent}}$ (named-entity coverage)—operate on each candidate. Each module is min–max normalized per query. The semantic module uses the same frozen BGE-large-en-v1.5 document embeddings and computes cosine similarity against $\mathrm{Normalize}(W_{\mathrm{BGE}}\boldsymbol{q}_{\mathrm{rep}})$. The lexical tokenizer lowercases, replaces ASCII punctuation with spaces, and then splits on whitespace, without stop-word removal or stemming. For the query and current N-document candidate pool, let $n = N + 1$ and let $\mathrm{df}(t)$ count rows containing term t. The lexical module uses raw term counts and $\mathrm{idf}(t) = \log[(1 + n)/(1 + \mathrm{df}(t))] + 1$, followed by cosine similarity. This is pool-local IDF: it discriminates terms within the current candidate set without a corpus-wide term-frequency table. A globally rare term common in this pool can consequently receive a low local weight; the score is not a measure of corpus-level rarity.

For the entity module, let $E_q(x)$ and $E_d(x)$ be sets of distinct, nonempty spaCy `en_core_web_sm` entity-surface strings. Query surfaces are lowercased and stripped; document surfaces are case-folded and stripped. This implementation asymmetry is immaterial for ASCII letters but can change non-ASCII matches; both operations are stated here to preserve the evaluated protocol. All spaCy entity types are retained, repeated surfaces count once, and matching uses the normalized surface string

without a type constraint. The raw score is

$$S_{\text{ent}}^{\text{raw}}(q,d) = \begin{cases} |E_q(q) \cap E_d(d)|/|E_q(q)|, & |E_q(q)| > 0, \\ 0, & |E_q(q)| = 0. \end{cases} \tag{15}$$

Thus, when the query contains no named entities, all raw entity scores are zero. A module whose scores are constant assigns a normalized value of 0.5 to every candidate. Exact MADS ties are resolved by document ID, and CCEF's stable sort preserves that order for exact fused-score ties. The primary configuration combines scores with fixed equal weights $w_j = \frac{1}{3}$; Appendix A.32 reports the corresponding combination study. Inter-module dispersion supplies CCEF's agreement signal, and MADS retains the top 100 candidates.

Score fusion (CCEF). CCEF fuses the three normalized module scores with an agreement multiplier:

$$s_{\text{fused}}(d_i) = \big(\textstyle\sum_j w_j \tilde{s}_{ij}\big)[\alpha + (1-\alpha)\text{agr}_i], \tag{16}$$

with $\alpha = 0.5$ and $\text{agr}_i = \text{clip}\big(1 - \sigma_i/(\mu_i + \epsilon), 0, 1\big)$, where $\mu_i = \text{mean}_j(\tilde{s}_{ij})$ and σ_i is the standard deviation across the three scores, using divisor 3 (`ddof=0`). The bounded multiplier lies in $[\alpha, 1]$ and equals 1 under perfect agreement. Documents below $\tau = 0.30$ are discarded, and CCEF retains the five highest-scoring threshold survivors. The primary policy selects exclusively from this survivor set and terminates the query if fewer than five remain (fail-closed selection). All primary runs retain at least five survivors.

Two ranking controls examine the dispersion estimator and threshold policy. The sample-dispersion control replaces $\sigma_i^{\text{pop}} = [\frac{1}{3}\sum_j(\tilde{s}_{ij} - \mu_i)^2]^{1/2}$ with $\sigma_i^{\text{samp}} = [\frac{1}{2}\sum_j(\tilde{s}_{ij} - \mu_i)^2]^{1/2}$ in the agreement multiplier. The score-backfill control first keeps all threshold survivors in descending fused-score order; if fewer than five remain, it appends the highest-fused-score discarded documents until the set contains five. Backfilled documents remain marked as below threshold for diagnostics. The primary configuration uses population dispersion and fail-closed selection; independently trained results for both alternatives appear in Appendix A.12.

Phase-II selected-document score surrogate. Retrieval and final top-five membership remain hard. For a selected document d_i, let $\widehat{\boldsymbol{q}}$ be the normalized projected QR state, $\widehat{\boldsymbol{v}}_i$ its normalized frozen BGE vector, and let $m_{\text{sem}}^{\text{stored}}$ and $\Delta_{\text{sem}}^{\text{stored}}$ be the semantic minimum and span recorded by the corresponding hard MADS pass. Its live normalized semantic coordinate is

$$\tilde{s}_{i,\text{sem}}^{\text{live}} = \begin{cases} (\langle \widehat{\boldsymbol{q}}, \widehat{\boldsymbol{v}}_i \rangle - m_{\text{sem}}^{\text{stored}})/\Delta_{\text{sem}}^{\text{stored}}, & \Delta_{\text{sem}}^{\text{stored}} > \epsilon, \\ 0.5, & \text{otherwise.} \end{cases} \tag{17}$$

Let $\boldsymbol{a}_i^{\text{live}}$ combine the stored normalized lexical and entity coordinates with this live semantic coordinate. Reapplying Eq. 16 with the configured dispersion estimator—SD_{pop} for the primary model and SD_{samp} only for the sample-dispersion control—yields the live fused score s_i^{live}. The score passed to the final fidelity blend and differentiable ordering surrogate is

$$s_i^{\text{ST}} = s_i^{\text{stored}} + s_i^{\text{live}} - \text{sg}(s_i^{\text{live}}) + I_i^{(2)}[z_i - \text{sg}(z_i)], \tag{18}$$

where $I_i^{(2)} = 1$ only when d_i was introduced by the second retrieval and $z_i = \langle \widehat{\boldsymbol{q}}_{\text{fb}}, \widehat{\boldsymbol{v}}_i \rangle$ is its feedback cosine. Equation 18 equals the stored fused score in the forward pass. In the backward pass, the live fused term provides the selected-document semantic path to QR, while the feedback term, which is zero in the forward pass, provides a path through P_{fb} and the first-pass compressed memories. ACR's score inputs remain detached.

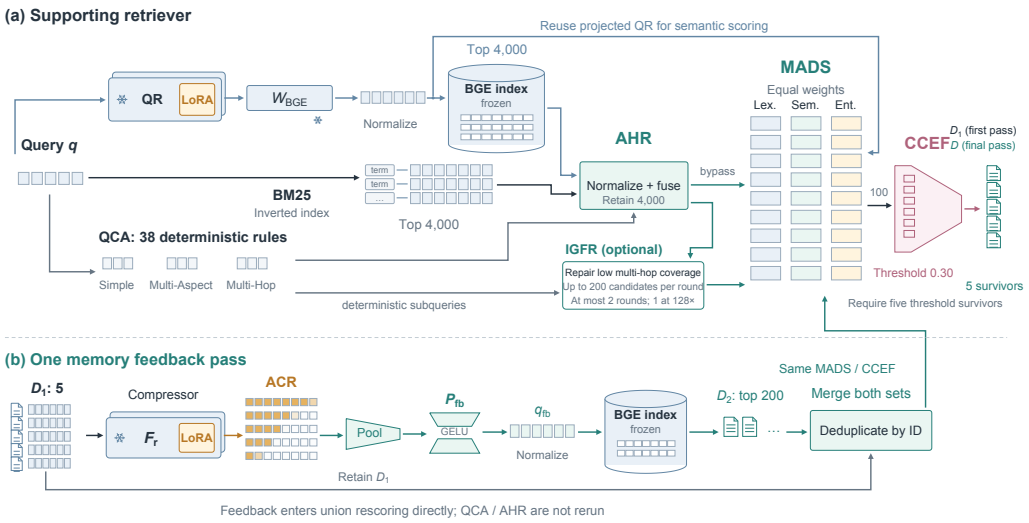

Figure A2: Supporting retrieval and one memory-feedback pass. The initial passages and 200 feedback candidates are merged, deduplicated, and rescored to select the final five passages. Feature rows are schematic.

Execution algorithm. Algorithm 1 summarizes BIRC's native execution. The query-reasoner (QR) forward pass precedes candidate acquisition: the dense retriever and semantic scorer consume $\boldsymbol{q}_{\text{rep}}$, and the routing rules operate directly on the query text. Coverage repair uses template-derived subqueries, and the final fusion step combines the normalized candidate scores.

Algorithm 1 BIRC execution with one feedback update.

Require: Query q; corpus $\mathcal{C}$; trained QR, compressor, reader, and feedback projection P_{fb}.

$\boldsymbol{q}_{\text{rep}} \leftarrow \text{QR}(q)$ ▷ query representation
type, $c(q)$, $\{\text{sub-q}_k\} \leftarrow \text{QCA}(q)$ ▷ deterministic routing
$C_1^{\text{AHR}} \leftarrow \text{AHR}(\boldsymbol{q}_{\text{rep}}, \text{type}, c(q); \mathcal{C}), |C_1^{\text{AHR}}| = 4{,}000$ ▷ acquisition
$C_1^{\text{aug}} \leftarrow C_1^{\text{AHR}}$
if type $=$ Multi-Hop **and** $\text{cov}(q, C_1^{\text{AHR}}) < \gamma$ **then**
  $C_1^{\text{aug}} \leftarrow C_1^{\text{AHR}} \cup \text{IGFR}(\{\text{sub-q}_k\}, C_1^{\text{AHR}}; K_{\text{igfr}})$
end if
$C_1^{100} \leftarrow \text{MADS}(\boldsymbol{q}_{\text{rep}}, C_1^{\text{aug}}), |C_1^{100}| = 100$ ▷ candidate scoring
$D_1 \leftarrow \text{CCEF}(C_1^{100}; \tau, \alpha), |D_1| = 5$ ▷ score fusion
$\{\rho_i\} \leftarrow \text{ACR-scale}(s_{\text{fused}}(D_1))$ ▷ Eq. 3
$\{\boldsymbol{m}_i\} \leftarrow \text{ACR-Gate}(\text{Compressor}(D_1), \{\rho_i\})$ ▷ Eq. 4
$\boldsymbol{q}_{\text{fb}} \leftarrow \text{MTFRL}(\{\boldsymbol{m}_i\}, \{T_i\})$ ▷ Eq. 6
$D_2 \leftarrow \text{DenseSearch}(\boldsymbol{q}_{\text{fb}}, \mathcal{C}), |D_2| = 200$ ▷ feedback retrieval
$U \leftarrow \text{Dedup}_{\text{docID}}(D_1 \cup D_2)$
$U^{100} \leftarrow \text{MADS}(\boldsymbol{q}_{\text{rep}}, U), |U^{100}| = 100$
$D \leftarrow \text{CCEF}(U^{100}; \tau, \alpha), |D| = 5$
$\{\rho_i^{\text{final}}\} \leftarrow \text{ACR-scale}(s_{\text{fused}}(D))$ ▷ recompute on final D
$\{\boldsymbol{m}_i^{\text{final}}\} \leftarrow \text{ACR-Gate}(\text{Compressor}(D), \{\rho_i^{\text{final}}\})$
$\{f_i^{\text{final}}\} \leftarrow \text{MinMax}(-\text{SPE}(D; \{\boldsymbol{m}_i^{\text{final}}\}))$ ▷ CFRS score
$s_{\text{final}}(d_i) \leftarrow w_{\text{cfrs}} f_i^{\text{final}} + (1 - w_{\text{cfrs}}) s_{\text{fused}}(d_i), \quad \forall d_i \in D$
$\pi \leftarrow \text{StableSort}_{\downarrow}(s_{\text{final}}(D))$
$D \leftarrow \pi(D); \{\boldsymbol{m}_i^{\text{ordered}}\} \leftarrow \pi(\{\boldsymbol{m}_i^{\text{final}}\})$
$\widehat{a} \leftarrow \text{Reader}(q, \{\boldsymbol{m}_i^{\text{ordered}}\})$
return $\widehat{a}$

All intermediate candidate and memory states belong to the current query and are discarded after decoding. Union rescoring uses the original query representation to retain question relevance. This also creates a bottleneck: bridge-only passages found by feedback can be dropped before reading.

Since the surrogate applies only to selected documents, discarded feedback candidates receive no direct selected-score gradient; continuous feedback does not make corpus search or top-five membership differentiable.

Oracle candidate pools. For each question, the frozen BGE retriever supplies passages from its 100 highest-ranked distinct pages. The highest-ranked passage represents each page. Annotated support pages absent from this list are appended in annotation order, replacing the same number of lowest-ranked non-support pages. The resulting 100-page pool is shared by BIRC and CLaRa. Retained candidates keep their BGE order, and support annotations affect pool membership only; each system applies its usual scoring and selection to the resulting passages.

CFRS passage ordering. For each selected candidate, CFRS prepends the fixed instruction "Reconstruct the original passage from the memory tokens. Output only the reconstructed passage," inserts the ACR-gated memory states, and teacher-forces all non-special source tokens retained under the 768-token passage-content limit; structural and special tokens are excluded from the target. The squared probability error is the mean, across passage positions, of the squared ℓ_2 distance between the full next-token probability vector and the one-hot gold target. This reconstruction pass disables all LoRA adapters and keeps the base-decoder weights frozen; gradients through the inserted memories continue to reach the compressor adapter. CFRS normalizes the negative error into f_i and applies Eq. 8.

A.2 Retrieval-Space Projections

Pre-fitted query-space map. Before either training phase, we fit one bias-free linear map W_{BGE}: $\mathbb{R}^{d_h} \to \mathbb{R}^{1024}$ for each backbone using a fixed set of 50,000 aligned question–passage records. Each record contains a stable passage ID and canonical page URL. Alignment records and official test records have disjoint page URLs. The records preserve a fixed alignment among questions, passage IDs, and indexed BGE representations. Each artifact row supplies the question, passage text, passage ID, and page URL; ordered row identities must agree with the BGE target-embedding artifact. This defines the fitting input contract, but the source-dataset breakdown, split composition, and sampling proportions of these 50,000 records are not documented. URL disjointness alone does not recover that provenance or establish question-level disjointness. Exact reconstruction of the fitting set therefore requires its original sample manifest. During this fitting stage, the freshly instantiated QR LoRA has a zero-initialized LoRA-B branch and is therefore functionally inactive. Let $\boldsymbol{h}_{\mathrm{base}}(q)$ denote the resulting frozen-base-decoder final-layer state at the last non-padding query position, with a maximum query length of 256 tokens. We minimize $\mathcal{L}_{\mathrm{align}} = \mathbb{E}_{(q,p)}[1 - \cos(W_{\mathrm{BGE}}\boldsymbol{h}_{\mathrm{base}}(q), \mathrm{BGE}(p))]$ for two epochs with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01), a learning rate of 5×10^{-4}, and a batch size of 128. The resulting per-backbone map is frozen and reused across seeds and compression ratios. Phase I then trains the compressor adapter; after Phase I, that adapter is copied into the Phase-II query-reasoner adapter as described in Section 3, without refitting W_{BGE}. The projected Phase-II QR state supplies the dense-query input used by AHR and MADS.

Trainable feedback projection. The two-layer map

$$P_{\mathrm{fb}}(\boldsymbol{x}) = W_2\,\mathrm{GELU}(W_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{19}$$

uses $4096 \to 2048 \to 1024$ for Mistral-7B/Llama-3-8B and $5120 \to 2560 \to 1024$ for Qwen2.5-14B. It maps the pooled memory state to the feedback query used for the second dense search. Including biases, these projections contain $10{,}488{,}832$ and $15{,}732{,}224$ parameters, respectively. The projection is initialized to match the pre-fitted W_{BGE} map: paired rows of W_1 are initialized as $[W_{\mathrm{BGE}}; -W_{\mathrm{BGE}}]$, W_2 as $[I, -I]$ in the corresponding paired coordinates, and the biases and any extra Qwen intermediate coordinates are initialized to zero. The identity $\mathrm{GELU}(z) - \mathrm{GELU}(-z) = z$ gives $P_{\mathrm{fb}}(\boldsymbol{x}) = W_{\mathrm{BGE}}\boldsymbol{x}$ at initialization.

Analytical Phase-II trainable-parameter counts. Under the configured rank-16 target modules, each of BIRC's query-reasoner, compressor, and reader `q_proj` adapters contains $4{,}194{,}304$ parameters for Mistral-7B and Llama-3-8B, and $7{,}864{,}320$ for Qwen2.5-14B. Including P_{fb}, the corresponding Phase-II totals are $23{,}071{,}744$ and $39{,}325{,}184$. CLaRa updates its query-reasoner and reader all-linear adapters in Phase II; each contains $41{,}943{,}040$ parameters for Mistral-7B/Llama-3-8B and

68,812,800 for Qwen2.5-14B, giving totals of 83,886,080 and 137,625,600, respectively. These counts follow from the configured module shapes. The native-system comparison matches the data, seeds, and training schedule while retaining each method's adapter placement and trainable-parameter count.

Phase II objective. Phase II uses the two terms in Eq. 11, with $\lambda = 0.10$. In the CFRS branch, let $e_i = \mathrm{SPE}(d_i)$ and $f_i^{\mathrm{raw}} = -e_i$, and define $\Delta = \max_j f_j^{\mathrm{raw}} - \min_j f_j^{\mathrm{raw}} + \epsilon$, so that $f_i = (f_i^{\mathrm{raw}} - \min_j f_j^{\mathrm{raw}})/\Delta$. Both $\min_j f_j^{\mathrm{raw}}$ and Δ are treated as detached normalization constants. Then

$$\frac{\partial f_i}{\partial e_i} = -\Delta^{-1}. \tag{20}$$

Consequently, the QA objective backpropagates through the per-document CFRS error while the ACR gate scales remain detached. Computing SPE adds one reconstruction-only forward pass per final candidate and reuses the compressed memory from the main computation. The five final candidates therefore require five such passes per training example. Each pass scores at most 768 target tokens in logit chunks of 128. Base-decoder parameters remain frozen, and all LoRA adapters are disabled within these reconstruction passes; CFRS therefore adds no decoder-side trainable parameters or optimizer state beyond the adapters already trained through the other Phase-II paths.

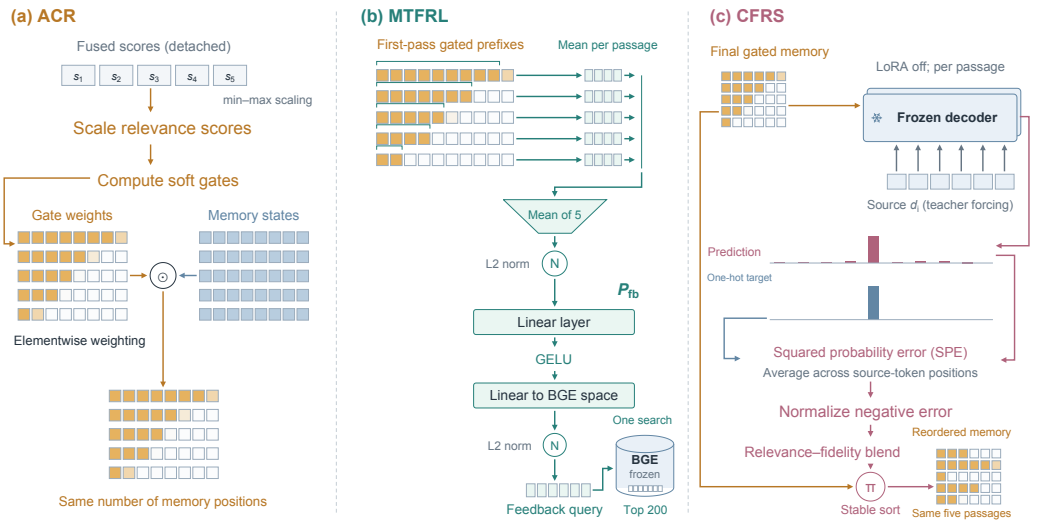

Figure A3: ACR gating, MTFRL pooling and projection, and CFRS ordering. Gating preserves tensor length; CFRS reorders the same five passage memories. Shading, probability bars, and the permutation are schematic.

A.3 QCA Rules and Routing Diagnostics

QCA declarative rule inventory. Tables A73 and A74 list the 38 surface-level rules used by QCA, organized by category. Appendix A.1 defines $c(q)$, specifies the routing precedence, and describes how $c(q)$ controls AHR. All string matching is case-insensitive; named-entity recognition (NER) uses spaCy (Honnibal et al., 2020) `en_core_web_sm`. All reported QCA results use this NER model. Rules within each group are ordered by descending weight. The weights are fixed coefficients in the declarative rule inventory, not learned model parameters. No weight-fitting protocol is documented, so their numerical precision should not be interpreted as calibrated confidence or an optimized routing policy.

Subquestion generation for IGFR. For queries classified as Multi-Hop, QCA additionally produces deterministic subquestions using templates for bridge connectives, possessive nesting, and "who-action" patterns (rules H02–H12 and H14 in Table A73). H01 and H13 use the first two named-entity spans in surface order as a deterministic fallback. A bridge template first asks for the

intermediate entity (e.g., "Where is Microsoft's headquarters?") and then instantiates the remaining predicate with the resolved bridge (e.g., "What is the tallest mountain in Washington state?"). During IGFR, bridge entities are resolved online from the accumulated retrieved passages. The implementation extracts the configured spaCy entity types, normalizes surfaces by case folding and stripping, excludes surfaces already in the query, and counts each remaining surface once per passage. It chooses the surface occurring in the most passages, breaking ties lexicographically, and substitutes it into the next `{BRIDGE}` template. If no surface is available, that template is consumed without issuing a query. This is a frequency heuristic, not relation extraction or entity linking, and uses no LLM generation.

Routing check. Against a 1,000-query reference panel, QCA achieves a weighted F1 score of 84.12%. Reference labels are used only to score the routing predictions after QCA has run; the rules, $c(q)$, and downstream checkpoints are label-independent. Table A1 reports downstream CEM after partitioning examples by the Full BIRC route assigned before the route-only intervention.

Table A1 compares Full QCA routing with a route-label-only Simple control. Both variants use the same checkpoints and examples. For the control, QCA first produces its complete surface record and then, before AHR, replaces only `question_type` with Simple. The control retains $c(q)$, the matched-rule trace, entity and hop counts, and template-derived subquestions. All downstream stages remain otherwise unchanged; AHR and IGFR operate on the modified record. Both variants are partitioned by the Full BIRC surface route computed before intervention.

Relative to the route-label-only Simple control, Full QCA routing yields CEM gains of 0.64 pp for Multi-Hop, 0.24 pp for Multi-Aspect, and 0.01 pp for Simple queries. Across all four benchmarks, Full QCA routing and the route-label-only Simple control have Avg CEM scores of 41.30 and 40.99, respectively, a difference of $+0.31$ pp.

This route-only control is distinct from the complete $-$QCA intervention in Table A50. The route-only control changes only the consumed route label, whereas $-$QCA removes the route, $c(q)$, and template-derived subquestions and disables AHR's type-specific weighting.

Table A1: Route-label control (CEM%, Normal, $16\times$; five-seed means). Simple is forced after QCA and before AHR; other QCA fields remain unchanged. Both variants are grouped by the original Full BIRC route. Avg weights benchmarks equally; bold marks route–benchmark pairwise maxima, including ties.

Variant	CR	Type	NQ	HotpotQA	MuSiQue	2Wiki	Avg
Full QCA route	16×	Simple	**62.39**	**51.06**	**14.97**	**55.76**	**46.05**
Full QCA route	16×	Multi-Hop	**55.20**	**41.95**	**10.66**	**44.05**	**37.97**
Full QCA route	16×	Multi-Aspect	**56.64**	**47.01**	**10.17**	**47.73**	**40.39**
Full QCA route	16×	Overall	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**
Route forced to Simple	16×	Simple	62.38	51.05	**14.97**	55.75	46.04
Route forced to Simple	16×	Multi-Hop	54.24	41.37	10.19	43.51	37.33
Route forced to Simple	16×	Multi-Aspect	56.34	46.77	10.00	47.48	40.15
Route forced to Simple	16×	Overall	58.35	45.86	11.49	48.26	40.99

A.4 Additional Controlled Experiments

The main controlled comparisons (§5.1) examine interface updates through fixed-checkpoint interventions and independent retraining. Two additional evaluations examine the first-pass direct-context endpoint and the choice of MuSiQue training partition.

First-pass direct-context endpoint. This condition runs the first QCA–AHR–IGFR–MADS–CCEF pass and feeds its five CCEF survivors directly into Mistral-7B. It omits compression, MTFRL, and CFRS and compares the first-pass text pipeline with Full BIRC. Appendix A.26 gives the full protocol and results.

MuSiQue partition control. A separate control retrains both systems on the standard 19,938-example MuSiQue partition rather than the augmented 168,745-example partition. Table A42 reports the complete comparison.

A.5 CEM Results under Normal and Oracle Retrieval

Under Normal retrieval, BIRC's Avg CEM exceeds CLaRa's at every tested ratio, with the largest mean difference, $+2.45$ pp, at $16\times$. Under Oracle retrieval, the largest difference is $+3.18$ pp at $16\times$ (Figure A1). All $128\times$ BIRC results use $K_{\text{igfr}} = 1$, as specified in Appendix A.28.

Table A2: Compression CEM% under Normal (N)/Oracle (O) retrieval. BIRC/CLaRa: five-seed means; other rows: single runs from Table 5 of He et al. (2025), retaining its normalization, CR labels, and averages. Paired tests and bold pairwise maxima concern only same-CR BIRC/CLaRa rows. BIRC uses $K_{\text{igfr}} = 1$ at $128\times$ (Appendix A.28).

Model	CR	Ret.	NQ	HotpotQA	MuSiQue	2Wiki	Avg
Normal Retrieval							
AutoCompressor	1×	N	17.24	14.61	3.81	19.89	13.89
Mistral-7B	1×	N	54.58	42.94	8.94	44.24	37.67
LLMLingua-2	4×	N	47.53	37.05	9.02	44.35	34.49
PISCO	16×	N	54.39	41.94	10.09	44.88	37.83
CLaRa	4×	N	57.18	45.24	10.40	47.08	39.98
CLaRa	16×	N	55.77	43.87	10.25	45.52	38.85
CLaRa	32×	N	54.79	42.55	9.72	44.53	37.90
CLaRa	64×	N	53.91	**42.72**	9.88	44.10	37.65
CLaRa	128×	N	53.06	41.14	9.61	**43.19**	36.75
BIRC	4×	N	**59.86**	**47.35**	**12.19**	**49.20**	**42.15**
BIRC	16×	N	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**
BIRC	32×	N	**56.88**	**44.09**	**10.02**	**45.61**	**39.15**
BIRC	64×	N	**55.34**	42.61	**10.08**	**44.69**	**38.18**
BIRC	128×	N	**54.12**	**41.72**	**9.63**	42.33	**36.95**
Oracle Retrieval							
Mistral-7B	1×	O	71.64	70.77	45.72	68.83	64.24
LLMLingua-2	4×	O	63.99	52.42	27.47	53.92	49.45
PISCO	16×	O	73.44	66.53	33.80	60.45	58.55
CLaRa	4×	O	76.38	73.96	46.14	70.64	66.78
CLaRa	16×	O	75.37	70.93	43.06	66.31	63.92
CLaRa	32×	O	73.06	68.35	38.54	64.17	61.03
CLaRa	64×	O	71.42	64.25	35.12	**60.89**	57.92
CLaRa	128×	O	68.84	61.93	31.12	**58.59**	55.12
BIRC	4×	O	**79.18**	**76.96**	**47.53**	**73.33**	**69.25**
BIRC	16×	O	**78.18**	**74.27**	**46.01**	**69.94**	**67.10**
BIRC	32×	O	**75.39**	**69.51**	**40.27**	**65.23**	**62.60**
BIRC	64×	O	**72.45**	**65.91**	**36.79**	60.45	**58.90**
BIRC	128×	O	**69.49**	**62.26**	**31.90**	58.35	**55.50**

A.6 Semantic Fidelity Under Compression

Table A3 reports passage-reconstruction fidelity under each compressor's native reconstruction protocol. BIRC uses fixed panels of 2,000 seen and 2,000 unseen passages; pages in the unseen panel are excluded from training by URL, and panel membership is fixed across ratios and seeds. Given the latent memory, we prompt the decoder with the fixed CFRS reconstruction instruction, decode greedily for at most 128 tokens, and score the output against the source text retained under the 768-token passage-content limit. BERTScore F1, ROUGE-1, ROUGE-L, and named-entity recall are computed per passage before aggregation.

The CLaRa rows report the Mistral-7B values and Seen/Unseen definitions from Table 16 of its information-preservation evaluation (He et al., 2025), rounded to one decimal place. That evaluation

uses CLaRa's native paraphrasing prompt and passage panels. Each block therefore describes compression-ratio trends under that method's own reconstruction protocol. For task-level comparisons, Table 3 gives native-pipeline endpoints, whereas Table 2 includes the shared-candidate control. The unseen panels score higher than the seen panels, but these are different passage sets, not a paired generalization contrast. Their length, entity density, and predictability are not matched. In particular, the 128-token reconstruction cap can affect scores against sources of different lengths. These are possible explanations, not measured causes; the aggregate scores alone do not establish superior reconstruction on unseen data.

Table A3: Semantic fidelity (%): BS = BERTScore F1 (Zhang et al., 2020), R1/RL = ROUGE-1/L, Ent = entity recall. BIRC uses 2,000 passages per split and five-seed means; CLaRa values are from its published evaluation. Unseen passages exclude compressor training data. Panels, prompts, and objectives are method-specific.

Model	**CR**	**Seen**				**Unseen**			
		BS	R1	RL	Ent	BS	R1	RL	Ent
(a) CLaRa native protocol; within-CLaRa trends only									
CLaRa	4×	90.7	55.9	40.1	54.8	91.5	59.7	44.1	60.0
CLaRa	16×	90.6	56.1	40.3	54.8	91.4	60.0	44.1	59.9
CLaRa	32×	90.6	56.2	40.1	53.9	91.4	60.3	44.0	59.3
CLaRa	128×	89.8	54.1	36.6	47.8	91.0	59.6	42.5	55.8
(b) BIRC native protocol; within-BIRC trends only									
BIRC	4×	91.8	57.9	42.3	57.4	92.6	61.9	45.7	63.3
BIRC	16×	91.6	57.4	42.0	56.8	92.4	61.6	45.4	62.8
BIRC	32×	91.3	56.8	41.7	56.2	92.2	61.2	45.1	62.2
BIRC	128×	89.2	55.7	35.8	49.9	91.6	60.5	43.1	57.4

A.7 Complete Native-Pipeline QA Results

Under each system's native retrieval pipeline, BIRC's five-seed Avg F1 is higher than that of our five-seed CLaRa baseline at every ratio from $4\times$ to $128\times$. The literature-reference DRO row (Shi et al., 2026) has a single-run Avg F1 of 41.96% under the setting reported in CLaRa's comparison table. Under Oracle retrieval at $4\times$, the BIRC–CLaRa Avg F1 difference is 4.16 pp. Across Normal-retrieval ratios, BIRC's Avg F1 advantage ranges from $+3.81$ to $+6.11$ pp.

Table A4: Native-pipeline F1/EM% under Normal (N)/Oracle (O) retrieval. BIRC/CLaRa: five-seed means; bold marks same-CR pairwise maxima. Single-run In-ctx RAG (Ram et al., 2023), RECOMP (Xu et al., 2024), GenGround (Shi et al., 2024), RetRobust (Yoran et al., 2024), and DRO (Shi et al., 2026) rows follow He et al. (2025) and are excluded from paired tests. *Variable compression. BIRC uses $K_{\text{igfr}} = 1$ at 128× (Appendix A.28).

Model	CR	Ret.	NQ		HotpotQA		MuSiQue		2Wiki		Avg F1	Avg EM
			F1	EM	F1	EM	F1	EM	F1	EM		
Normal Retrieval												
In-ctx RAG	1×	N	44.69	38.07	41.27	37.14	20.11	16.78	41.02	38.51	36.77	32.62
RECOMP	var.*	N	42.67	37.47	42.72	38.72	24.96	17.34	38.26	32.17	37.15	31.43
GenGround	1×	N	42.31	40.60	44.71	41.27	24.36	20.77	42.58	39.61	38.49	35.56
RetRobust	1×	N	43.82	37.03	40.54	35.59	18.16	18.11	39.11	38.65	35.41	32.34
DRO	1×	N	51.01	42.41	47.87	40.37	25.32	21.36	43.65	42.12	41.96	36.56
CLaRa	4×	N	48.34	38.16	46.07	35.12	17.55	8.11	47.32	43.11	39.82	31.12
CLaRa	16×	N	47.94	37.84	45.85	35.01	17.68	8.23	44.33	40.09	38.95	30.29
CLaRa	32×	N	47.28	37.22	44.28	33.95	16.72	7.85	42.82	38.52	37.77	29.38
CLaRa	64×	N	46.41	36.51	44.65	34.18	16.08	7.52	40.57	36.34	36.93	28.64
CLaRa	128×	N	45.54	35.63	42.29	31.92	14.50	6.68	**38.31**	**34.11**	35.16	27.08
BIRC	4×	N	**56.18**	**46.16**	**54.23**	**42.57**	**22.37**	**12.18**	**50.94**	**47.61**	**45.93**	**37.13**
BIRC	16×	N	**55.19**	**44.83**	**52.71**	**41.29**	**21.30**	**11.48**	**48.80**	**44.52**	**44.50**	**35.53**
BIRC	32×	N	**54.87**	**44.26**	**49.63**	**39.18**	**19.27**	**9.79**	**47.11**	**42.89**	**42.72**	**34.03**
BIRC	64×	N	**53.06**	**42.57**	**49.87**	**39.36**	**19.13**	**10.01**	**44.82**	**40.58**	**41.72**	**33.13**
BIRC	128×	N	**51.83**	**41.89**	**47.96**	**37.69**	**18.07**	**9.11**	38.02	33.39	**38.97**	**30.52**
Oracle Retrieval												
CLaRa	4×	O	75.78	67.64	69.81	56.92	33.28	22.42	74.02	69.74	63.22	54.18
CLaRa	16×	O	71.68	63.29	68.57	55.38	30.86	20.56	67.98	62.64	59.77	50.47
CLaRa	32×	O	69.88	65.17	**66.08**	**53.41**	28.94	18.45	64.52	59.32	57.36	49.09
CLaRa	64×	O	67.57	62.08	63.61	51.43	27.16	16.73	61.61	56.28	54.99	46.63
CLaRa	128×	O	65.00	58.94	61.36	**49.63**	23.65	14.97	58.43	53.17	52.11	44.18
BIRC	4×	O	**79.91**	**72.39**	**75.48**	**62.25**	**36.82**	**25.26**	**77.31**	**73.15**	**67.38**	**58.26**
BIRC	16×	O	**76.63**	**68.56**	**73.29**	**59.85**	**34.06**	**23.16**	**73.26**	**68.57**	**64.31**	**55.04**
BIRC	32×	O	**73.89**	**66.18**	65.88	53.28	**33.21**	**22.51**	**69.48**	**64.89**	**60.62**	**51.72**
BIRC	64×	O	**71.58**	**63.88**	**65.20**	**52.76**	**28.62**	**18.70**	**68.18**	**63.18**	**58.40**	**49.63**
BIRC	128×	O	**69.96**	**61.49**	**61.48**	49.24	**27.27**	**17.96**	**64.02**	**58.98**	**55.68**	**46.92**

A.8 Unified Rescoring at 16×

Table A5 applies the same EM/CEM/F1 evaluator, using one canonical reference answer per example, to outputs from BIRC, SARA (Jin et al., 2026), SeleCom (Liu et al., 2026), OSCAR (Louis et al., 2026), CLaRa (He et al., 2025), and COCOM (Rau et al., 2025).

Table A5: Unified single-reference EM/CEM/F1% (16×, Normal); Avg weights four benchmarks equally. BIRC/CLaRa use five-seed means; other methods use complete output panels with native configurations and seed counts. Δ Avg F1: BIRC minus method (pp). Bold marks metric maxima.

Method	NQ EM/CEM/F1	HotpotQA EM/CEM/F1	MuSiQue EM/CEM/F1	2Wiki EM/CEM/F1	Avg EM	Avg CEM	Avg F1	Δ Avg F1 (pp)
BIRC	**44.83/58.73/55.19**	**41.29/46.17/52.71**	**11.48/11.74/21.30**	**44.52/48.56/48.80**	**35.53**	**41.30**	**44.50**	0.00
SARA	42.27/57.59/52.77	39.12/45.80/50.17	10.29/11.39/20.51	43.86/47.59/48.07	33.89	40.59	42.88	+1.62
SeleCom	41.58/57.17/51.75	38.24/45.18/49.38	9.94/11.11/19.92	43.25/47.17/47.35	33.25	40.16	42.10	+2.40
OSCAR	40.04/56.35/50.69	36.95/44.81/48.40	9.37/10.80/19.13	42.54/46.56/46.69	32.23	39.63	41.23	+3.27
CLaRa	37.84/55.77/47.94	35.01/43.87/45.85	8.23/10.25/17.68	40.09/45.52/44.33	30.29	38.85	38.95	+5.55
COCOM	36.38/53.11/46.91	33.09/41.53/43.91	7.53/9.64/16.96	38.22/43.88/42.17	28.81	37.04	37.49	+7.01

A.9 Training-Schedule-Matched Native-System Comparisons

The unified-rescoring table above retains each released system's native training protocol and seed count. As a separate control, we retrain SARA (Jin et al., 2026), SeleCom (Liu et al., 2026), OSCAR (Louis et al., 2026), and CLaRa (He et al., 2025) with the same `mistralai/`

`Mistral-7B-Instruct-v0.2` base model and tokenizer, Phase-II QA examples, Phase-II schedule, nominal 16× ratio, five seeds, and the single-reference answer evaluator used for BIRC. This control fixes those training and evaluation variables while each method retains its native retrieval path, compression mechanism, objective, candidate pool, and processing steps.

Table A6: Training-schedule-matched comparison (16×, Normal; five-seed means). Retrieval, candidate pools, and latency boundaries remain method-specific. Δ Avg F1: BIRC minus method (pp). Bold marks task-score maxima.

Method	NQ F1	Hotpot F1	MuSiQue F1	2Wiki F1	Avg F1	Avg CEM	Sec./query	BIRC−method
BIRC	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**	**41.30**	6.62	0.00
SARA	53.43	51.07	20.37	48.33	43.30	40.75	6.41	1.20
SeleCom	52.47	50.13	19.96	47.64	42.55	40.25	6.24	1.95
OSCAR	51.84	49.37	19.26	47.33	41.95	39.75	6.11	2.55
CLaRa	47.94	45.85	17.68	44.33	38.95	38.85	5.83	5.55

Table A7: Schedule-matched five-seed Avg F1. Marginal 95% CIs: Student's t, 4 df; paired BIRC-minus-method CIs: query-level bootstrap. p-values are Holm-adjusted over four contrasts.

Method	42	123	456	789	2024	Mean	SD	Marginal 95% CI	BIRC−method 95% CI	p_{Holm}
BIRC	44.76	44.37	44.16	44.62	44.59	44.50	0.24	[44.21,44.79]	—	—
SARA	43.94	42.92	42.54	43.67	43.43	43.30	0.57	[42.60,44.00]	[0.63,1.77]	0.003
SeleCom	43.31	42.07	42.01	42.90	42.46	42.55	0.55	[41.86,43.24]	[1.39,2.51]	< 0.001
OSCAR	42.66	41.52	41.21	42.41	41.95	41.95	0.60	[41.20,42.70]	[1.92,3.18]	< 0.001
CLaRa	39.10	38.84	38.70	39.22	38.89	38.95	0.21	[38.69,39.21]	[5.33,5.77]	< 0.001

BIRC exceeds SARA, the highest-scoring competitor in this training-schedule-matched protocol, by 1.20 pp in Avg F1 (paired query-level bootstrap 95% CI $[0.63, 1.77]$; Holm-adjusted $p = 0.003$). These results use the matched five-seed training schedule; the native-training output panels are reported separately in Table A5.

A.10 Pretraining Data Composition

A no-pretraining baseline, four cumulative pretraining-mixture configurations, and a standalone paraphrase-only control are compared over five seeds in Table A8 (CEM%, Oracle, nominal CR = 32×). The Full row exceeds SQ+CQ+Para by 2.06 pp in Avg CEM, with differences of +3.06 pp on MuSiQue and +1.34 pp on 2Wiki. The SimpleQA row exceeds No pretraining by 3.40 pp, and SQ+ComplexQA exceeds SimpleQA by 1.71 pp. Paraphrase-only training reaches 59.80% Avg CEM, while the full four-category mixture attains the highest Avg CEM, 62.60%. This experiment uses 32× Oracle retrieval; the primary operating point in the main text is 16× Normal retrieval.

Table A8: Pretraining mixtures (CEM%, Oracle, nominal 32×; five-seed means). SQ through Full add data categories cumulatively; Paraphrase-only is independent. SQ/CQ/Para/Ent denote SimpleQA/ComplexQA/Paraphrase/Entity-Augmented. Bold marks metric maxima.

Data	NQ	HotpotQA	MuSiQue	2Wiki	Avg
No pretraining	70.97	62.44	29.35	58.16	55.23
+SimpleQA	72.42	66.12	34.14	61.85	58.63
SQ+ComplexQA	74.01	68.38	35.81	63.16	60.34
SQ+CQ+Para	73.60	67.44	37.21	63.89	60.54
Full (SQ+CQ+Para+Ent)	**75.39**	**69.51**	**40.27**	**65.23**	**62.60**
Paraphrase-only (standalone)	73.07	67.53	36.02	62.56	59.80

A.11 QA Hidden-State Alignment Loss: Presence and λ Sensitivity

Table A9 compares training with and without the squared-ℓ_2 QA hidden-state alignment loss $\mathcal{L}_{\mathrm{HSA}}$ (Eq. 13) at 32× and 128× under Oracle retrieval. The λ sweep below uses 32× Oracle retrieval. This loss is distinct from the per-document CFRS statistic. At 32×, the difference in Avg CEM between training with and without the loss is 0.77 pp. The largest per-benchmark difference is on HotpotQA (+1.33 pp), followed by MuSiQue (+0.79 pp). At 128×, the corresponding Avg CEM difference is 0.89 pp, including a +1.05 pp difference on HotpotQA. Avg CEM is higher with the loss at both compression ratios.

Table A9: HSA loss ablation (CEM%, Oracle; five-seed means), with other training settings fixed. Averages use unrounded values; bold marks within-CR maxima.

Condition	CR	NQ	HotpotQA	MuSiQue	2Wiki	Avg CEM
Without $\mathcal{L}_{\mathrm{HSA}}$	32×	74.99	68.18	39.48	64.68	61.83
With $\mathcal{L}_{\mathrm{HSA}}$	32×	**75.39**	**69.51**	**40.27**	**65.23**	**62.60**
Without $\mathcal{L}_{\mathrm{HSA}}$	128×	68.72	61.21	31.01	57.50	54.61
With $\mathcal{L}_{\mathrm{HSA}}$	128×	**69.49**	**62.26**	**31.90**	**58.35**	**55.50**

Sensitivity to λ. The weight λ is swept over $\{0, 0.01, 0.05, 0.1, 0.2, 0.5\}$ at 32× Oracle in Table A10. The primary setting, $\lambda = 0.1$, has the highest Avg CEM (62.60%), and $\lambda = 0.01$ is 0.24 pp above the $\lambda = 0$ baseline. The tested settings $\lambda \in \{0.05, 0.10, 0.20\}$ lie within 0.20 pp of the default. The full sweep spans 0.77 pp, and the three central settings span 0.19 pp. We use $\lambda = 0.1$ in all main experiments.

Table A10: HSA weight sensitivity (CEM%, Oracle, 32×; five-seed means), with other settings fixed. $\lambda = 0$ matches Table A9. Δ Avg CEM: row minus primary $\lambda = 0.10$ (pp). Bold marks the primary row.

λ	NQ	HotpotQA	MuSiQue	2Wiki	Avg CEM	Δ Avg CEM (pp)
0.00	74.99	68.18	39.48	64.68	61.83	−0.77
0.01	75.15	68.57	39.68	64.88	62.07	−0.53
0.05	75.32	69.27	39.94	65.12	62.41	−0.19
0.10	**75.39**	**69.51**	**40.27**	**65.23**	**62.60**	+0.00
0.20	75.09	69.41	39.89	65.41	62.45	−0.15
0.50	74.31	68.92	39.07	65.67	61.99	−0.61

A.12 HSA Reduction, Gradient Routing, and CCEF Selection Policies

We vary two aspects of HSA: its reduction over hidden dimensions and the states that receive its gradients. For training example b, let $\boldsymbol{\delta}_b = \overline{\boldsymbol{h}}_{\mathcal{M},b} - \overline{\boldsymbol{h}}_{\mathcal{P}_{\mathrm{QA}},b}$ and let d_h be the hidden dimension. The two reductions are

$$\ell_b^{\mathrm{sum}} = \sum_{k=1}^{d_h} \delta_{bk}^2, \qquad \ell_b^{\mathrm{mean}} = \frac{1}{d_h} \sum_{k=1}^{d_h} \delta_{bk}^2. \tag{21}$$

For gradient routing, we either retain gradients through both pooled states (bidirectional), detach the memory mean, or detach the query–answer mean before computing $\boldsymbol{\delta}_b$. The primary objective in Eq. 13 uses sum reduction and bidirectional gradients. We also examine the two CCEF alternatives defined in Appendix A.1. All runs use Mistral-7B, Normal retrieval, 16× compression, and the paired seeds 42/123/456/789/2024. Each control is independently trained with the same Phase-I initialization and Phase-II schedule as the primary model. HSA controls retain the primary CCEF policy, and CCEF controls retain the primary sum-reduced, bidirectional HSA objective; all other settings are fixed.

In Tables A11 and A12, "mean" denotes the hidden-coordinate mean ℓ^{mean}, whereas "sum" denotes the unnormalized ℓ^{sum} used by the primary model. "Stop-M" and "Stop-QA" stop the HSA gradient

through the pooled memory state and pooled query–answer state, respectively; "bi." leaves both pooled states attached. Each CCEF control changes one choice: the dispersion estimator or the policy for fewer than five threshold survivors.

Table A11: HSA/CCEF variants (CEM/F1%, Normal, Mistral-7B, 16×; five-seed means). Each variant changes only its named choices. Δ: variant minus primary (pp). Bold marks metric maxima, including ties.

Variant	NQ	HotpotQA	MuSiQue	2Wiki	Avg	Δ Avg
Primary: sum, bi.; pop., fail-closed	58.73/55.19	46.17/52.71	11.74/21.30	**48.56/48.80**	41.30/44.50	$+0.00/+0.00$
HSA: mean, bi.	57.81/54.58	46.25/52.76	10.93/20.58	48.01/48.52	40.75/44.11	$-0.55/-0.39$
HSA: sum, Stop-M	58.28/54.85	**46.29**/52.80	11.36/21.01	48.31/48.62	41.06/44.32	$-0.24/-0.18$
HSA: sum, Stop-QA	**58.98/55.37**	46.05/52.61	**11.92/21.45**	48.53/48.77	**41.37/44.55**	$+0.07/+0.05$
HSA: mean, Stop-M	57.70/54.42	46.23/52.75	10.82/20.62	48.01/48.41	40.69/44.05	$-0.61/-0.45$
HSA: mean, Stop-QA	57.95/54.65	46.26/52.78	11.08/20.81	48.15/48.52	40.86/44.19	$-0.44/-0.31$
CCEF: sample dispersion	58.49/55.03	46.28/**52.81**	11.55/21.16	48.48/48.72	41.20/44.43	$-0.10/-0.07$
CCEF: score backfill	58.80/55.24	46.13/52.68	11.79/21.33	**48.56**/48.79	41.32/44.51	$+0.02/+0.01$

Table A12: Per-seed Avg CEM/F1% for Table A11; Mean averages five paired seeds. Bold marks metric maxima within each column.

Variant	42	123	456	789	2024	Mean
Primary	41.09/44.76	41.52/44.37	41.31/44.16	41.35/44.62	**41.22**/44.59	41.30/44.50
HSA: mean, bi.	40.21/44.09	41.56/44.43	40.35/43.45	41.04/44.44	40.58/44.14	40.75/44.11
HSA: sum, Stop-M	40.61/44.40	41.61/44.49	40.94/43.85	41.13/44.52	41.00/44.34	41.06/44.32
HSA: sum, Stop-QA	**41.40/44.98**	41.34/44.21	**41.55/44.35**	**41.44**/44.59	41.11/**44.62**	**41.37/44.55**
HSA: mean, Stop-M	40.07/43.97	41.58/44.42	40.23/43.32	40.93/44.31	40.63/44.23	40.69/44.05
HSA: mean, Stop-QA	40.38/44.21	41.62/44.45	40.49/43.55	41.06/44.45	40.74/44.29	40.86/44.19
CCEF: sample dispersion	40.81/44.58	**41.67/44.50**	41.09/44.00	41.41/44.60	41.01/44.47	41.20/44.43
CCEF: score backfill	41.16/44.80	41.47/44.34	41.34/44.14	41.41/**44.67**	41.21/44.60	41.32/44.51

With $\lambda = 0.1$ held fixed, dimension-wise averaging scales both the HSA value and its direct gradient by $1/d_h$ relative to the primary sum reduction; it therefore changes the effective regularization strength. The primary bidirectional sum reduction exceeds the bidirectional mean reduction by 0.55 points in Avg CEM and 0.39 points in Avg F1. Under sum reduction, Stop-M changes the two averages by $-0.24/-0.18$ points and Stop-QA by $+0.07/+0.05$ points. Stop-QA has a positive paired difference for three of the five seeds in each metric.

Using sample rather than population dispersion increases the across-scorer standard deviation by a factor of $\sqrt{3/2}$ whenever dispersion is nonzero, which increases the disagreement penalty before clipping. In the sample-dispersion control, Avg CEM/F1 is $0.10/0.07$ points lower than in the primary configuration. Score backfill changes Avg CEM/F1 by $+0.02/+0.01$ points, with both positive and negative differences across paired seeds. These controls characterize sensitivity around the primary configuration of sum-reduced bidirectional HSA, population dispersion, and selection from threshold survivors.

A.13 Reader-Only vs. Joint Phase-II Adapter Training

Reader-only fine-tuning and joint Phase-II training of the QR, compressor, and reader LoRA adapters plus P_{fb} are compared in Table A13. Under Normal retrieval, the Avg CEM difference is $+1.47$ pp at 32× and $+1.27$ pp at 128×; under Oracle retrieval at 32×, it is $+2.25$ pp. NQ has the largest per-benchmark difference in all three settings.

Table A13: Fine-tuning scope (CEM%, Normal/Oracle; five-seed means): reader LoRA only versus joint QR/compressor/reader LoRA and P_{fb} (default). Bold marks within-condition maxima.

Scope	CR	Ret.	NQ	HotpotQA	MuSiQue	2Wiki	Avg
Reader-only	32×	N	54.63	42.86	9.60	43.64	37.68
Joint adapters	32×	N	**56.88**	**44.09**	**10.02**	**45.61**	**39.15**
Reader-only	128×	N	52.15	40.13	8.99	41.44	35.68
Joint adapters	128×	N	**54.12**	**41.72**	**9.63**	**42.33**	**36.95**
Reader-only	32×	O	72.94	67.11	38.48	62.86	60.35
Joint adapters	32×	O	**75.39**	**69.51**	**40.27**	**65.23**	**62.60**

A.14 MADS Scorer Combination Study

Table A14 reports MADS scorer combinations under Oracle retrieval at 32×; Appendix A.32 gives the corresponding Normal-retrieval study at 16×. The full Lex+Sem+Ent configuration has the highest value in every displayed Oracle column and the highest Avg F1 under Normal retrieval; its largest Oracle margin over Lex+Sem is 1.01 pp on MuSiQue. All rows are fixed-checkpoint test-time scoring interventions averaged over five seeds.

Table A14: Fixed-checkpoint MADS combinations (CEM%, Oracle, 32×; five-seed means). Bold marks benchmark maxima.

Config	Lex	Sem	Ent	NQ	HotpotQA	MuSiQue	2Wiki
Lex only	✓	×	×	73.84	68.26	39.23	63.52
Sem only	×	✓	×	74.97	68.91	39.53	64.65
Lex+Sem	✓	✓	×	74.89	69.24	39.26	64.93
Lex+Sem+Ent	✓	✓	✓	**75.39**	**69.51**	**40.27**	**65.23**

A.15 IGFR Iteration, CFRS Weight, and AHR Weight Sensitivity

Tables A15–A17 report fixed-checkpoint sensitivity to the IGFR iteration count, CFRS weight, and AHR weight pairs around the reported primary configuration. The benchmark evaluation grids were also used to select K_{igfr} and w_{cfrs}. At 16×, the selected values 2 and 0.30 maximize measured Avg CEM in these grids. These are exploratory, evaluation-selected results, not an independent assessment of a development-selected configuration. The following development procedure applies only to the three named parameter families.

Development-set hyperparameter selection. For the CCEF agreement floor, ACR sigmoid sharpness, and ACR scale bounds, we evaluated the full Cartesian product on a 500-query held-out development split disjoint from training and evaluation. The search used Normal retrieval at 16× and Avg CEM computed before rounding as its selection metric: $\alpha \in \{0.3, 0.5, 0.7\}$, $\beta \in \{5, 10, 20\}$, and $\rho_{\min} \in \{0.1, 0.25, 0.4\}$, with $\rho_{\max} = 1.0$. Exact ties are resolved in favor of the first tuple in the listed order. This procedure selected $\alpha = 0.5$, $\beta = 10$, and $(\rho_{\min}, \rho_{\max}) = (0.25, 1.0)$. Except where an intervention explicitly replaces or bypasses one of these parameters, all main and ablation experiments use the selected values.

IGFR iterations. The primary $K_{\text{igfr}} = 2$ row (41.30% Avg CEM) exceeds the $K_{\text{igfr}} = 3$ and $K_{\text{igfr}} = 4$ rows by 0.23 and 0.48 pp, respectively.

Table A15: IGFR iteration sensitivity (CEM%, Normal, 16×; five-seed means). Bold marks metric maxima and primary K_{igfr}.

K_{igfr}	NQ	HotpotQA	MuSiQue	2Wiki	Avg
1	58.20	45.79	11.31	48.18	40.87
2	**58.73**	46.17	**11.74**	**48.56**	**41.30**
3	58.55	46.02	11.49	48.22	41.07
4	57.81	**46.25**	11.26	47.96	40.82

CFRS blending weight. Table A16 sweeps w_{cfrs}. At the primary weight $w_{\text{cfrs}} = 0.30$, NQ CEM is 1.25 pp higher than at $w_{\text{cfrs}} = 0$.

Table A16: CFRS weight sensitivity (CEM%, Normal, 16×; five-seed means). $w_{\text{cfrs}} = 0$ removes fidelity reranking; bold marks the primary row and maxima.

w_{cfrs}	NQ	Hotpot	MuSiQue	2Wiki	Avg
0.00	57.48	45.42	11.29	48.01	40.55
0.10	58.15	45.75	11.35	48.47	40.93
0.20	58.60	46.00	11.55	48.49	41.16
0.30	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**
0.40	58.68	45.73	11.27	48.40	41.02
0.50	57.82	45.28	11.08	48.50	40.67

AHR defaults. Table A17 varies candidate $(w_{\text{bm25}}, w_{\text{dense}})$ pairs for each QCA category (five-seed Avg CEM under Normal retrieval at 16×). The Simple, Multi-Hop, and Multi-Aspect ranges are 0.32, 0.77, and 0.22 pp, respectively. The listed $\boldsymbol{w}_{\text{type}}$ endpoints retain a lexical component, and $c(q)$ interpolates each endpoint with $(0.50, 0.50)$ as specified in §A.1.

Table A17: AHR weight sensitivity (Avg CEM%, Normal, 16×; five-seed means). One category varies; others stay fixed. Avg weights four benchmarks equally. Δ: row minus the category default. Bold marks each default row.

Category	$(w_{\text{bm25}}, w_{\text{dense}})$	Avg CEM	Δ Avg CEM (pp)
Simple	(0.90, 0.10)	41.14	−0.16
	(0.75, 0.25)	**41.30**	0.00
	(0.50, 0.50)	40.98	−0.32
Multi-Hop	(0.10, 0.90)	40.92	−0.38
	(0.25, 0.75)	**41.30**	0.00
	(0.50, 0.50)	40.53	−0.77
Multi-Aspect	(0.50, 0.50)	41.08	−0.22
	(0.40, 0.60)	41.19	−0.11
	(0.30, 0.70)	**41.30**	0.00
	(0.20, 0.80)	41.24	−0.06

A.16 CFRS Fidelity Criterion: SPE versus Cross-Entropy

The default CFRS fidelity criterion is squared probability error (SPE). We compare SPE with token cross-entropy, $-|\mathcal{T}|^{-1} \sum_{t \in \mathcal{T}} \log p(y_t \mid y_{<t}, \boldsymbol{m})$, computed with the same teacher-forced instruction format and target-token mask. In both cases, we negate the passage-level loss, min–max normalize it over the selected five passages, and blend the resulting fidelity score with fused retrieval relevance using the same 0.3 weight. In the fixed-checkpoint comparison, the fidelity score is the sole evaluation-time change; the retrained comparison trains each configuration with its corresponding criterion.

Table A18: CFRS fidelity controls (16×, Normal; five-seed means). Avg F1 CIs: 95%, Student's t, 4 df. None removes fidelity reranking; bold marks SPE and task-score maxima within each suite. Holm correction covers two SPE-minus-cross-entropy contrasts.

Suite	Fidelity score	Avg CEM	Avg F1	Sec./query	Avg F1 95% CI
Fixed checkpoint	None	40.55	43.70	6.55	[43.29,44.11]
	Token cross-entropy	41.02	44.20	6.62	[43.84,44.56]
	SPE	**41.30**	**44.50**	6.62	[44.21,44.79]
Independent retraining	Token cross-entropy	41.05	44.29	6.62	[43.94,44.64]
	SPE	**41.20**	**44.47**	6.62	[44.12,44.82]

Within the fixed-checkpoint suite, the SPE-minus-cross-entropy difference in Avg F1 is 0.30 points (95% paired query-level bootstrap CI $[0.21, 0.39]$; Holm-adjusted $p < 0.001$). After independent retraining, the corresponding difference is 0.18 points (95% paired query-level bootstrap CI $[0.11, 0.25]$; Holm-adjusted $p = 0.001$). SPE yields higher Avg F1 than token cross-entropy under both protocols.

A.17 Dataset and Pretraining Statistics

Tables A19 and A20 summarize the pretraining corpus and the QA training and evaluation splits.

Phase II training subsample protocol. The four benchmarks contain $485{,}006$ training examples in total. Phase II uses a *benchmark-balanced* subsample of $38{,}400$ examples per epoch ($9{,}600$ from each benchmark), drawn uniformly without replacement and reshuffled at every epoch. Every run uses the epoch-level sampling seeds $\{42, 123, 456, 789, 2024\}$, corresponding to the five epochs. This schedule is distinct from the training-run seed assigned to the checkpoint. For Mistral-7B and Llama-3-8B, five epochs at an effective batch size of 32 give $192{,}000$ example views and $6{,}000$ optimizer steps; Qwen2.5-14B uses the same $192{,}000$ views with an effective batch size of 16, giving $12{,}000$ optimizer steps. The balanced schedule assigns equal sampling weight to each benchmark rather than weighting benchmarks by their raw split sizes. The derived MuSiQue partition contributes $9{,}600$ examples per epoch from its $168{,}745$-example pool, an inclusion rate of 5.69% per epoch. The corresponding NQ rate is 16.38% ($9{,}600$ of $58{,}622$).

Phase-I pretraining mixture. The four Phase-I strata total $7{,}808{,}465$ training pairs. Their names denote functional strata in the prepared corpus rather than public benchmark releases. Each input row contains a source document, a separate reconstruction target, unique source and target identifiers, and a canonical page URL. The data preparation canonicalizes evaluation URLs, removes exact page matches, and verifies the post-exclusion count for each stratum. Each training example is a pair (d, y): the frozen base decoder predicts the target sequence y from the compressed memory $\mathcal{F}_r(d)$ alone, without the raw document or a category-specific instruction. Each stratum uses the fixed system instruction "You are a helpful assistant. Reconstruct the background from its memory tokens and output only the paraphrase." For Mistral-7B and Llama-3-8B, each Phase-I epoch deterministically shuffles all $7{,}808{,}465$ pairs and consumes $7{,}808{,}448$ examples in complete effective batches of 32, dropping the 17-example shuffled tail. Qwen2.5-14B consumes $7{,}808{,}464$ examples in complete effective batches of 16 and drops the one-example shuffled tail. SimpleQA, ComplexQA, Paraphrase, and Entity-Augmented are internal functional mixture labels; Table A19 reports their aggregate composition.

Evaluation benchmarks. At the benchmark level, we describe NQ as single-hop, HotpotQA and 2Wiki as two-hop, and MuSiQue as two-to-four-hop. These labels summarize benchmark construction; individual questions can differ in the reasoning they require. MuSiQue has the smallest evaluation set (2,417 examples), whereas 2Wiki has the largest (12,576 examples).

MuSiQue training partition. The versioned $168{,}745$-row derived MuSiQue source comprises original, subquestion, partial-chain, and entity-variant families. The partition is $8.46\times$ the answerable-only subset of $19{,}938$ examples. Evaluation uses the standard 2,417-example split without augmentation.

Table A19: Phase-I sample counts by stratum. Para.: Paraphrase; Ent-Aug.: Entity-Augmented.

	SimpleQA	**ComplexQA**	**Para.**	**Ent-Aug.**
Samples	2,000,000	2,000,000	1,966,291	1,842,174

Table A20: Processed benchmark counts. †HotpotQA: two documents; ‡MuSiQue: 2–4 hops. §MuSiQue uses 168,745 augmented training examples versus 19,938 standard answerable examples (Appendix A.43). Other training counts include preprocessing and page-URL deduplication.

Dataset	**Train**	**Eval**	**Hops**
NQ	58,622	6,489	1
HotpotQA	90,185	7,384	2†
MuSiQue	168,745§	2,417	2–4‡
2Wiki	167,454	12,576	2

A.18 Cross-Reader Results: Llama-3-8B

We repeat the core experiments on Llama-3-8B-Instruct (Grattafiori et al., 2024) under the same two-phase protocol used for Mistral-7B: Phase I runs for 3 epochs, and Phase II uses a learning rate of 2×10^{-4}, a 500-step warmup, 5 epochs, an effective batch size of 32, and LoRA rank $r_{\mathrm{LoRA}}=16$. The MTFRL projection head uses the same dimensions as the corresponding Mistral-7B head ($4096\to2048\to1024$). The training data match the Mistral-7B configuration, including the derived MuSiQue training partition (Appendix A.17). We use the same five seeds $\{42, 123, 456, 789, 2024\}$. Tables A21 and A22 report Normal retrieval. In the endpoint tables, boldface identifies the larger value within each same-CR BIRC–CLaRa pair.

BIRC has higher Avg CEM than CLaRa at all tested compression ratios; the largest mean difference is $+2.92$ pp at 16× (vs. $+2.45$ pp on Mistral-7B). At the same 16× point, the F1 difference is $+6.16$ pp (47.60% vs. 41.44%), compared with $+5.55$ pp on Mistral-7B. Across the full ratio range, the Avg CEM advantage is $+1.12$–$+2.92$ pp.

Table A21: Llama-3-8B-Instruct native-pipeline CEM% (Normal; five-seed means). Bold marks same-CR BIRC/CLaRa pairwise maxima.

Model	**CR**	**NQ**	**Hotpot**	**MuSiQue**	**2Wiki**	**Avg**
CLaRa	4×	59.84	47.62	13.05	49.53	42.51
CLaRa	16×	58.21	46.58	12.83	48.10	41.43
CLaRa	32×	57.30	45.12	11.95	46.81	40.30
CLaRa	64×	56.40	45.38	11.24	45.60	39.66
CLaRa	128×	55.50	43.20	10.40	**45.60**	38.68
BIRC	4×	**63.06**	**50.31**	**14.96**	**52.39**	**45.18**
BIRC	16×	**61.99**	**49.16**	**14.45**	**51.80**	**44.35**
BIRC	32×	**60.63**	**47.93**	**13.36**	**50.29**	**43.05**
BIRC	64×	**59.40**	**46.60**	**13.53**	**48.94**	**42.12**
BIRC	128×	**57.92**	**44.79**	**11.29**	45.19	**39.80**

Table A22: Llama-3-8B-Instruct native-pipeline F1% (Normal; five-seed means). Bold marks same-CR BIRC/CLaRa pairwise maxima.

Model	CR	NQ	Hotpot	MuSiQue	2Wiki	Avg
CLaRa	4×	51.24	49.20	20.78	48.34	42.39
CLaRa	16×	50.37	48.41	20.18	46.80	41.44
CLaRa	32×	49.49	46.80	19.32	45.09	40.18
CLaRa	64×	48.58	47.08	18.77	43.35	39.45
CLaRa	128×	47.72	44.73	17.76	**41.74**	37.99
BIRC	4×	**60.02**	**57.61**	**25.63**	**53.50**	**49.19**
BIRC	16×	**58.63**	**55.76**	**24.07**	**51.94**	**47.60**
BIRC	32×	**57.68**	**53.45**	**22.65**	**50.68**	**46.12**
BIRC	64×	**56.11**	**53.81**	**21.34**	**49.42**	**45.17**
BIRC	128×	**54.51**	**49.72**	**19.56**	40.07	**40.97**

Under Oracle retrieval (Table A23), the Avg CEM differences are $+3.32$ pp at 16× and $+0.78$ pp at 128×. The 16× difference is 0.14 pp larger than the corresponding Mistral-7B difference of $+3.18$ pp.

Table A23: Llama-3-8B-Instruct native-pipeline CEM% (Oracle; five-seed means). Bold marks same-CR BIRC/CLaRa pairwise maxima.

Model	CR	NQ	Hotpot	MuSiQue	2Wiki	Avg
CLaRa	16×	77.83	73.18	45.56	69.34	66.48
BIRC	16×	**80.66**	**76.84**	**48.69**	**73.01**	**69.80**
CLaRa	128×	71.19	63.85	33.47	59.79	57.08
BIRC	128×	**72.03**	**64.49**	**34.12**	**60.80**	**57.86**

A.19 Interface-Update Sensitivity Across Compression Ratios

We apply fixed-checkpoint evaluation-time interventions to trained Full BIRC models. In the CFRS-free condition, the CCEF order is passed directly to the reader. The Uniform-gate condition replaces the score-conditioned scale with the fixed value $\rho = 0.625$ while retaining the gating operation. The No-feedback-retrieval condition omits the second retrieval round. In the All-updates-off condition, CFRS and feedback retrieval are disabled, and ACR is bypassed by passing all $K_i^{(0)}$ base states directly to the reader.

Table A24 extends the fixed-checkpoint interface-update analysis to all compression ratios and reports full per-benchmark results for every intervention.

Table A24: Fixed-checkpoint interface interventions (CEM%, Normal; five-seed means). At 128×, $K_{\text{igfr}} = 1$. Bold marks within-CR maxima and rows containing a maximum.

CR	Variant	NQ	HotpotQA	MuSiQue	2Wiki	Avg
4×	**Full**	**59.86**	**47.35**	**12.19**	**49.20**	**42.15**
	CFRS-free	59.00	46.71	11.94	48.67	41.58
	Uniform gate	59.47	46.97	11.59	48.45	41.62
	No feedback retrieval	59.52	46.43	11.80	48.77	41.63
	All updates off	57.31	45.25	10.92	47.12	40.15
16×	**Full**	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**
	CFRS-free	57.48	45.42	11.29	48.01	40.55
	Uniform gate	58.08	45.02	10.79	47.51	40.35
	No feedback retrieval	58.13	44.72	11.09	47.74	40.42
	All updates off	56.03	43.82	10.34	46.21	39.10
32×	**Full**	**56.88**	**44.09**	**10.02**	**45.61**	**39.15**
	CFRS-free	56.15	43.65	9.74	45.26	38.70
	Uniform gate	56.52	43.40	9.57	44.99	38.62
	No feedback retrieval	56.52	43.20	9.85	45.31	38.72
	All updates off	54.86	42.11	9.14	43.69	37.45
64×	**Full**	**55.34**	**42.61**	**10.08**	**44.69**	**38.18**
	CFRS-free	54.89	42.25	9.96	44.42	37.88
	Uniform gate	55.05	42.20	9.94	44.49	37.92
	No feedback retrieval	55.12	41.83	9.98	44.47	37.85
	All updates off	54.00	41.21	9.35	43.36	36.98
128×	**Full**	**54.12**	41.72	9.63	42.33	36.95
	CFRS-free	53.53	41.36	9.48	42.03	36.60
	Uniform gate	52.89	**42.17**	**10.08**	**44.26**	**37.35**
	No feedback retrieval	53.88	41.44	9.29	42.35	36.74
	All updates off	52.60	41.08	9.73	43.23	36.66

Table A25 also reports the per-benchmark CEM and F1 values at the primary 16× operating point.

Table A25: Fixed-checkpoint interface interventions (16×, Normal; five-seed mean CEM/F1%). Bold marks column maxima.

Variant	CEM%					F1%				
	NQ	Hot.	MuSi.	2Wiki	Avg	NQ	Hot.	MuSi.	2Wiki	Avg
Full	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**
CFRS-free	57.48	45.42	11.29	48.01	40.55	54.12	51.87	20.34	48.47	43.70
Uniform gate	58.08	45.02	10.79	47.51	40.35	54.47	51.73	19.74	48.26	43.55
No feedback retrieval	58.13	44.72	11.09	47.74	40.42	54.58	51.37	20.14	48.39	43.62
All updates off	56.03	43.82	10.34	46.21	39.10	51.93	49.17	18.34	45.76	41.30

A.20 Independently Retrained Interface-Update Configurations

Table A26 compares five independently trained configurations at 16× under Normal retrieval. The Full checkpoints in this suite (41.20 Avg CEM and 44.47 Avg F1) are distinct from the primary native-pipeline checkpoints (41.30 and 44.50). Each variant is trained with its stated forward path. CFRS-free uses the retrieval-score order. Uniform gate assigns all five passages the same score-independent scale, $\rho = \min\{1, 108/\sum_i K_i^{(0)}\}$, while retaining all base memory positions. The MTFRL control replaces memory feedback with the original query for the second retrieval. All updates off combines these three substitutions. In contrast, the fixed-checkpoint All-updates-off control bypasses gating and omits the second retrieval entirely. Per-seed Avg F1 values and intervals appear in Table A27; the Avg CEM means for this suite appear in Table A26.

Table A26: Retrained interface controls (16×, Normal; five-seed means). Δ Avg F1: Full minus variant (pp). Full is bold.

	CEM%					F1%					Δ Avg F1 (pp)
Configuration	**NQ**	**Hotpot**	**MuSiQue**	**2Wiki**	**Avg**	**NQ**	**Hotpot**	**MuSiQue**	**2Wiki**	**Avg**	**Full−variant**
Full	**58.62**	**46.03**	**11.69**	**48.46**	**41.20**	**55.11**	**52.57**	**21.06**	**49.14**	**44.47**	—
CFRS-free	57.92	45.52	11.33	47.83	40.65	54.61	52.15	20.71	48.21	43.92	+0.55
Uniform gate	58.07	45.12	10.93	47.80	40.48	54.65	51.92	20.08	48.35	43.75	+0.72
Original-query second retrieval	58.03	44.97	11.13	48.03	40.54	54.71	51.73	20.39	48.37	43.80	+0.67
All updates off	55.93	44.07	10.58	46.62	39.30	52.29	49.27	18.53	45.99	41.52	+2.95

Table A27: Retrained controls: per-seed Avg F1 and sample SD. Marginal and paired Full-minus-variant intervals are 95% CIs (Student's t, 4 df).

Configuration	**42**	**123**	**456**	**789**	**2024**	**Mean**	**SD**	**95% CI (t, 4 df)**	**Full−variant 95% CI**
Full	44.78	44.29	44.07	44.61	44.60	**44.47**	0.29	[44.12, 44.82]	—
CFRS-free	44.27	43.62	43.54	44.08	44.09	43.92	0.32	[43.52, 44.32]	[+0.47, +0.63]
Uniform gate	44.12	43.39	43.31	43.96	43.97	43.75	0.37	[43.29, 44.21]	[+0.58, +0.86]
Original-query second retrieval	44.15	43.46	43.41	44.01	43.97	43.80	0.34	[43.38, 44.22]	[+0.56, +0.78]
All updates off	42.06	41.17	41.02	41.67	41.68	41.52	0.42	[41.00, 42.04]	[+2.76, +3.14]

A.21 Conditional Component-Swap Protocol

For each training seed s, the swap evaluation uses independently trained BIRC and CLaRa checkpoints, denoted A_s and C_s, matched by seed, compression ratio, base-model release, tokenizer, memory geometry, and reader input format. The two directions exchange passage IDs, text, and scalar scores. Latent states are recomputed by the receiving system, whose compressor and reader come from the same checkpoint. We evaluate all four benchmarks over five paired training seeds and aggregate metrics from the resulting per-example interface records.

Table A28: Configurations in Appendices A.19, A.21, and A.39 (Mistral-7B, Normal, 16×; five-seed means). A_s/C_s: paired-seed BIRC/CLaRa checkpoints.

Configuration	**Retrieval**	**Compressor–reader**	**CFRS**	**ACR**	**MTFRL**	**Avg F1**
Full BIRC	A_s	A_s	✓	✓	✓	44.50
All updates off	A_s	A_s	×	×	×	41.30
BIRC-ret → CLaRa-stack (CFRS)	A_s	C_s	✓	×	×	43.30
BIRC-ret → CLaRa-stack (off)	A_s	C_s	×	×	×	41.24
CLaRa-ret → BIRC-stack	C_s	A_s	✓	✓	×	41.55
CLaRa	C_s	C_s	*N/A (native)*			38.95

BIRC retrieval to the CLaRa stack. A_s runs one QCA–AHR–IGFR–MADS–CCEF pass. For this component-swap evaluation, CCEF returns its top 20 ranked candidates before the usual five-passage cutoff. Their IDs, texts, and fused scores cross the interface. C_s recompresses the 20 passages and applies its native selector to retain five. The CFRS-off configuration passes all base memory states and performs neither CFRS nor a second search. For the CFRS-on endpoint, the 20-candidate input, five selected passages, and CLaRa stack remain fixed; CFRS blends teacher-forced reconstruction fidelity with the incoming CCEF score using the fixed 0.3 weight and reorders the selected passages. Thus, the two rows in Table A30 differ only in CFRS ordering. Appendix A.22 retrains CLaRa on the same BIRC top-20 interface.

CLaRa retrieval to the BIRC stack. C_s compresses the precomputed top-20 candidates returned by the frozen BGE retriever and applies its trained hard top-five selector. Its selected memory states are discarded; only the five passage IDs, texts, and selector cosine scores enter A_s. To place these

values on CFRS's $[0, 1]$ score scale, the interface applies

$$\widetilde{s}_i = \begin{cases} \dfrac{s_i - \min_j s_j}{\max_j s_j - \min_j s_j + 10^{-6}}, & \max_j s_j - \min_j s_j > 10^{-6}, \\ \frac{1}{2}, & \text{otherwise}, \end{cases} \tag{22}$$

within the selected five while preserving the selector's hard order for ties. A_s recompresses the passages, applies ACR using these scores, reranks them with CFRS in its own compressor–reader stack, and generates the answer. MTFRL is off, so the CLaRa retrieval pool is not expanded after the interface.

Table A29 reports the two directions together with the native-pipeline endpoints.

Table A29: Component swaps (F1%, Normal, 16×; five paired-seed means). Stack comprises the same checkpoint's compressor and reader. Bold marks maxima.

System	NQ	HotpotQA	MuSiQue	2Wiki	Avg F1
CLaRa (baseline)	47.94	45.85	17.68	44.33	38.95
CLaRa-ret → BIRC-stack (ACR+CFRS)	51.43	48.96	19.47	46.34	41.55
BIRC-ret → CLaRa-stack (CFRS)	53.47	51.28	20.43	48.02	43.30
BIRC (full)	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**

The BIRC-to-CLaRa swap reaches 41.24 Avg F1 with CFRS off and 43.30 with CFRS on (Table A30). Both rows reuse the same paired checkpoints, BIRC top-20 candidates, and CLaRa-selected five passages.

Table A30: CFRS in the BIRC-retrieval/CLaRa-stack hybrid (F1%, Normal, 16×; five paired-seed means). Retrieval outputs and checkpoint pairs are fixed; only ordering changes.

System	NQ	HotpotQA	MuSiQue	2Wiki	Avg F1
CLaRa (baseline)	47.94	45.85	17.68	44.33	38.95
BIRC-ret → CLaRa-stack (CFRS off)	51.73	49.08	18.23	45.92	41.24
BIRC-ret → CLaRa-stack (CFRS on)	53.47	51.28	20.43	48.02	43.30

A.22 Selector-Input Controls and Matched Factorial

Native CLaRa selects from the top 20 returned by the frozen BGE retriever. To compare candidate sources while holding the selector input at 20 items, we instead supply the 20 highest-ranked candidates produced by BIRC's first-pass scorer before its usual five-passage cutoff. Relative to native CLaRa, the test-time row changes only the selector's input list; the retrained row trains CLaRa's selector and reader on that interface, retaining its frozen compressor. The base model, reader input format, nominal ratio, Phase-II examples, training schedule, and five seeds match those used in the native CLaRa runs. This control fixes the selector input size at 20 while leaving upstream candidate acquisition method-specific.

Table A31: Twenty-candidate selector controls (16×, Normal; five-seed means). Middle rows pair BIRC's ranked top 20 with the CLaRa stack. Δ: Full-minus-row Avg F1 (pp). Full is bold.

System	NQ F1	Hotpot F1	MuSiQue F1	2Wiki F1	Avg F1	Avg CEM	Δ from Full
CLaRa, native BGE top-20 pool	47.94	45.85	17.68	44.33	38.95	38.85	5.55
BIRC top-20, test-time replacement	51.73	49.08	18.23	45.92	41.24	39.06	3.26
BIRC top-20, CLaRa retrained	52.31	49.43	18.61	45.97	41.58	39.43	2.92
Full BIRC	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**	**41.30**	**0.00**

Table A32: Selector controls: per-seed Avg F1, sample SD, and 95% CIs (Student's t, 4 df).

System	42	123	456	789	2024	Mean	SD	95% CI (t, 4 df)
CLaRa, native	39.10	38.84	38.70	39.22	38.89	38.95	0.21	[38.69,39.21]
BIRC top-20, test-time replacement	41.62	41.17	40.85	41.46	41.10	41.24	0.30	[40.86,41.62]
BIRC top-20, retrained	41.99	41.48	41.19	41.87	41.37	41.58	0.34	[41.16,42.00]
Full BIRC	44.76	44.37	44.16	44.62	44.59	44.50	0.24	[44.21,44.79]

The Avg F1 differences from native CLaRa are $+2.29$ points for test-time candidate replacement and $+2.63$ points after retraining; the corresponding 95% paired query-level bootstrap CIs are $[2.08, 2.49]$ and $[2.39, 2.86]$. Full BIRC's Avg F1 is 2.92 points higher than that of the retrained control ($[2.65, 3.20]$) and 3.26 points higher than that of the test-time replacement ($[3.05, 3.47]$). All four contrasts have Holm-adjusted $p < 0.001$ within this comparison family.

We also cross first-pass retrieval source with ACR+CFRS on the BIRC stack. MTFRL is disabled in every cell. The resulting factorial contrasts quantify retrieval-source and ACR+CFRS differences within this MTFRL-off BIRC stack. We report each factor's average conditional contrast, the difference between those contrasts, and their interaction.

Table A33: Matched 2×2 factorial ($16\times$, Normal, BIRC stack, MTFRL off). (a) Per-seed Avg F1%. (b) Five-seed contrasts (pp), with 95% CIs (Student's t, 4 df).

(a) Factorial cells

First-pass source	ACR+CFRS	42	123	456	789	2024	Mean
CLaRa	Off	39.22	38.71	38.64	39.39	39.29	39.05
BIRC	Off	41.63	41.02	40.88	41.55	41.42	41.30
CLaRa	On	41.84	41.15	41.02	41.89	41.85	41.55
BIRC	On	44.08	43.29	43.21	43.71	43.81	43.62

(b) Factorial contrasts

Contrast	Δ Avg F1	95% CI (t, 4 df)	Δ Avg CEM
First-pass retrieval (averaged)	2.16	[1.99,2.33]	0.26
ACR+CFRS (averaged)	2.41	[2.30,2.52]	1.26
ACR+CFRS minus retrieval	0.25	[0.09,0.41]	1.00
Retrieval × ACR+CFRS	-0.18	$[-0.31,-0.05]$	0.12

The averaged ACR+CFRS contrast exceeds the averaged first-pass retrieval contrast by 0.25 points in Avg F1. Each value averages its two conditional contrasts within the MTFRL-off BIRC stack. The -0.18-point interaction is the difference between the ACR+CFRS contrast under BIRC first-pass retrieval and the same contrast under CLaRa first-pass retrieval—a difference-in-differences. Both conditional ACR+CFRS contrasts are positive (2.32 and 2.50 points); the interaction reflects the slightly smaller ACR+CFRS increment under BIRC first-pass retrieval.

A.23 Shared Top-Five Candidates from Frozen BGE

At $16\times$, this control constructs one page-unique top-five set per question with the frozen BGE retriever and reuses the same document IDs, page URLs, passage text, and acquisition scores for BIRC and CLaRa throughout Phase II and evaluation. Both systems are trained independently for each paired seed. BIRC bypasses QCA, AHR, IGFR, MADS, and CCEF and disables MTFRL; it applies its compressor–reader stack, ACR, and CFRS to the fixed five. CLaRa applies its native compressor–reader stack and selector to those five, so selection preserves all five passage identities and determines their order. BIRC's five-seed Avg F1 is 41.35, compared with 38.45 for CLaRa. Table A34 gives the benchmark-level results, and Table A35 reports the five-seed summary.

Table A34: Shared frozen-BGE top-five comparison (16×, Normal; EM/CEM/F1%). Aggregates average five paired seeds and four benchmarks. Δ Avg F1: BIRC minus CLaRa (pp). Bold marks maxima.

System	NQ EM/CEM/F1	HotpotQA EM/CEM/F1	MuSiQue EM/CEM/F1	2Wiki EM/CEM/F1	Avg EM	Avg CEM	Avg F1	Δ Avg F1 (pp)
CLaRa	37.13/54.82/47.12	34.21/43.02/45.01	7.73/9.72/17.19	40.25/45.08/44.48	29.83	38.16	38.45	—
BIRC stack ACR+CFRS MTFRL off	**40.17/56.58/50.57**	**36.86/44.72/48.36**	**9.34/10.93/19.14**	**42.95/47.17/47.33**	**32.33**	**39.85**	**41.35**	**+2.90**

Table A35: Shared-candidate retraining: per-seed Avg F1, sample SD, and 95% CIs (Student's t, 4 df). The paired row is BIRC minus CLaRa.

Configuration	42	123	456	789	2024	Mean	SD	95% CI (t, 4 df)
CLaRa	37.93	38.74	38.35	38.61	38.62	38.45	0.32	$[38.05, 38.85]$
BIRC stack (ACR+CFRS; MTFRL off)	40.98	41.29	41.65	41.46	41.37	41.35	0.25	$[41.04, 41.66]$
BIRC−CLaRa	+3.05	+2.55	+3.30	+2.85	+2.75	+2.90	0.29	$[+2.54, +3.26]$

At 128×, we evaluate five fixed checkpoints per system on a page-unique frozen-BGE top-five candidate pool constructed once per question and shared across systems and seeds. We use `BAAI/bge-large-en-v1.5`, batches of eight questions, and greedy decoding with a 64-new-token limit. Averaged over five seeds, Avg F1 is 35.98% for BIRC and 34.88% for CLaRa.

Table A36: Evaluation-only frozen-BGE top-five control (128×, Normal; five-seed mean CEM/F1%). Δ: BIRC minus CLaRa. Bold marks pairwise maxima.

System	NQ CEM/F1	Hotpot CEM/F1	MuSiQue CEM/F1	2Wiki CEM/F1	Avg CEM	Avg F1
CLaRa	52.64/45.07	40.77/42.18	**9.57**/14.26	**43.22/38.01**	36.55	34.88
BIRC	**53.66/47.24**	**41.25/43.98**	9.38/**14.99**	42.91/37.71	**36.80**	**35.98**
Δ	+1.02/+2.17	+0.48/+1.80	−0.19/+0.73	−0.31/−0.30	+0.25	+1.10

Table A37: 128× shared-candidate control: per-seed Avg F1. Marginal 95% CIs use Student's t (4 df); the paired contrast uses a benchmark-stratified query bootstrap for its CI and p-value.

System	42	123	456	789	2024	Mean	SD	95% CI	p
CLaRa	35.05	34.65	34.72	35.12	34.86	34.88	0.20	[34.63,35.13]	—
BIRC	35.93	35.93	35.68	36.44	35.92	35.98	0.28	[35.63,36.33]	—
BIRC−CLaRa	+0.88	+1.28	+0.96	+1.32	+1.06	+1.10	0.19	[0.72,1.48]	< 0.001

The paired Avg F1 difference is $+1.10$ pp (95% benchmark-stratified paired query-level bootstrap CI $[0.72, 1.48]$; $p < 0.001$; this analysis tests the single Avg F1 contrast).

A.24 Fixed-Checkpoint Latent and Text-Space Feedback Substitutions

We compare MTFRL with two text-space feedback controls: RM3-style pseudo-relevance feedback (PRF) (Abdul-Jaleel et al., 2004) and a HyDE-style generated hypothesis (Gao et al., 2023a). Each feedback arm starts from the same BIRC first-pass state and issues one 200-candidate follow-up search over the frozen BGE index. All arms then use the same union rescoring, five-passage selection, CFRS ordering, and reader. The three feedback arms share the 200-candidate follow-up budget; their measured latencies differ. The None arm omits the follow-up search. All four conditions are evaluation-time interventions on the same MTFRL-trained checkpoints. Thus, the experiment compares MTFRL, RM3-style PRF, and HyDE-style feedback as test-time alternatives within a shared stack.

We use the following fixed settings. The RM3-style PRF arm uses all five nonempty first-pass documents, selects ten feedback terms, and interpolates the original and feedback query models with weight 0.5. Document probabilities are computed with a numerically stable softmax over first-pass scores. Term probabilities use maximum-likelihood frequencies after NFKC normalization, Unicode case folding, alphanumeric tokenization, removal of terms in a fixed list of 33 stopwords, and filtering to tokens comprising 2–30 Unicode code points; no stemming is applied. Feedback documents are ordered by decreasing retrieval score, increasing input rank, and Unicode code-point order of the document ID; terms are ordered by decreasing weight and then by Unicode code-point order. Retrieved candidates are ordered by decreasing cosine score and increasing corpus row index. The feedback query begins with the question with leading and trailing whitespace removed. Any retained feedback terms that do not already occur in the query are appended after a newline, in decreasing order of interpolated weight.

The HyDE arm generates one passage with `mistralai/Mistral-7B-Instruct-v0.2`, its tokenizer, and no task adapter. It uses the following prompt, which contains two blank-line separators: one before `Question:` and one before `Passage:`.

```
<s>[INST] Write one concise, factual passage that directly answers the
search question. Write only the passage: do not mention these
instructions, do not say that the passage is hypothetical, and do not
add citations.

Question: {question}

Passage: [/INST]
```

We replace `{question}` with the question after removing leading and trailing whitespace and disable automatic special-token insertion. Decoding is greedy, with one beam, no sampling, a single returned sequence, a 128-new-token limit, and termination at EOS. After leading and trailing whitespace is removed, the generated passage—without the instruction prompt—is encoded for retrieval. Both text arms encode their retrieval input with the same frozen `BAAI/bge-large-en-v1.5` encoder used for the corpus. Retrieval uses cosine similarity between ℓ_2-normalized float32 vectors and returns 200 candidates.

Table A38: Test-time feedback substitutions in MTFRL-trained checkpoints (16×, Normal; five-seed means). Task scores and post-feedback supporting-page R@5: %; batch-1 wall-clock latency: seconds/query. Δ sec./F1: row minus None (seconds/pp). Bold marks MTFRL and score maxima.

Feedback	Avg CEM	Avg F1	Final R@5	Sec./query	Δ sec.	Δ F1
None	40.42	43.62	61.20	6.36	0.00	0.00
RM3-style PRF	40.78	43.95	62.80	6.50	+0.14	+0.33
HyDE-style text	40.92	44.07	63.40	7.78	+1.42	+0.45
MTFRL	**41.30**	**44.50**	**64.50**	6.62	+0.26	**+0.88**

Table A39: Feedback controls: per-seed Avg F1. Marginal 95% CIs use Student's t (4 df); MTFRL contrasts use paired query bootstrap CIs with Holm-adjusted p-values across three contrasts.

(a) Per-seed Avg F1

Feedback	42	123	456	789	2024	Mean	SD	95% CI (t, 4 df)
None	44.08	43.29	43.21	43.71	43.81	43.62	0.36	[43.17,44.07]
RM3-style PRF	44.38	43.61	43.54	44.10	44.12	43.95	0.36	[43.50,44.40]
HyDE-style	44.48	43.75	43.62	44.25	44.25	44.07	0.37	[43.61,44.53]
MTFRL	44.76	44.37	44.16	44.62	44.59	44.50	0.24	[44.21,44.79]

(b) Paired MTFRL contrasts

Contrast	Δ Avg F1	95% paired query-level bootstrap CI	p_{Holm}
MTFRL−none	0.88	[0.66,1.10]	< 0.001
MTFRL−RM3-style PRF	0.55	[0.32,0.77]	0.002
MTFRL−HyDE-style	0.43	[0.20,0.65]	0.006

Within this fixed-checkpoint protocol, MTFRL attains the highest Avg F1 and Final R@5. Its Avg F1 advantages over RM3-style PRF and HyDE-style feedback are 0.55 and 0.43 pp, respectively; it is also 1.16 seconds faster than HyDE-style feedback.

A.25 Number of Test-Time Feedback Rounds

BIRC checkpoints are trained with a single feedback round. We evaluate depths $T \in \{0, 1, 2, 3\}$ with all parameters fixed. Initially, S_0 and U_0 are both the sequence of five first-pass candidates in rank order. For $t \geq 1$, let S_{t-1} be the current top five and U_{t-1} the accumulated candidate sequence. Round t recomputes the gated memories from S_{t-1}, forms the corresponding feedback query, retrieves 200 candidates, and appends previously unseen document IDs in retrieval order:

$$U_t = \mathrm{first}_{\mathrm{docID}}(U_{t-1} \,\|\, R_t), \qquad S_t = \mathrm{Top5}(\mathrm{Rescore}(U_t)) \,. \tag{23}$$

Deduplication spans the initial candidate set and all feedback rounds; the first-seen passage record and global position are retained. Score ties are broken by that first position and then by document ID. Prior memory states are discarded. After T rounds, CFRS reorders S_T once without changing its membership.

Table A40: Test-time MTFRL depth (16×, Normal; five-seed means). Δ F1 is relative to one round. Bold marks the default and score maxima.

Rounds	Avg CEM	Avg F1	Final R@5	Sec./query	Peak GB	Δ F1
0	40.42	43.62	61.20	6.36	21.10	-0.88
1	41.30	44.50	64.50	6.62	21.20	0.00
2	**41.43**	**44.68**	65.40	6.89	21.30	+0.18
3	41.36	44.61	**65.70**	7.17	21.40	+0.11

Relative to no feedback, the default one-round configuration records an Avg F1 difference of $+0.88$ pp. Applying the checkpoint trained with one feedback round for two or three rounds at test time gives Avg F1 differences of $+0.18$ and $+0.11$ pp relative to the default, with latency increases of 0.27 and 0.55 seconds per query, respectively. All rows use the same checkpoints; only the number of test-time feedback rounds changes.

A.26 First-Pass Direct-Context Endpoint

This evaluation-only condition executes one QCA–AHR–IGFR–MADS–CCEF pass and concatenates the five surviving passages directly into the Mistral-7B reader prompt. It omits compression, MTFRL, and CFRS while retaining the Phase II reader-adapter checkpoint. Results are averaged over five seeds. With the reader tokenizer, the rounded means are 590 tokens per passage and $2{,}950$ document tokens per query. The model's context limit is 32,768 tokens. We reserve 64 positions for generation; when a prompt exceeds the remaining budget, only the evidence tail is truncated, preserving the system instruction and question.

Table A41 reports this direct-context endpoint alongside the compressed systems. The endpoint uses the first-pass CCEF survivors, whereas Full BIRC selects its final five passages after feedback retrieval and union rescoring. This is therefore a comparison of complete evidence-processing paths.

The two endpoints have different benchmark profiles: Full BIRC has higher F1 on NQ, HotpotQA, and 2Wiki, while direct context has higher F1 on MuSiQue. Their four-benchmark Avg F1 scores are 44.50% and 45.40%, respectively. Direct context supplies the first-pass passages as text; Full BIRC supplies compressed memories of the final passages selected after feedback retrieval. The reported contrast therefore covers both evidence selection and reader input representation. Appendix A.33 separately evaluates gate-floor changes while holding the trained checkpoints fixed.

Latency and peak memory for this first-pass CCEF endpoint are not reported. Table A67's 8.43-second/32.4-GB direct-context profile uses frozen-BGE top-five passages and cannot be paired with this endpoint's 45.40 Avg F1. Establishing an accuracy–latency frontier against the stronger endpoint requires timing its own first-pass retrieval and reader invocation on the same panel and hardware.

Table A41: Fixed-reader, evaluation-only direct context versus CLaRa/Full BIRC (Normal; five-seed means, task metrics in %). Direct context reads five first-pass CCEF survivors without compression, MTFRL, or CFRS; compressed systems use 16×, and Full BIRC selects passages after feedback. Bold marks metric maxima.

System	CR	NQ		HotpotQA		MuSiQue		2Wiki		Avg F1	Avg CEM
		F1	EM	F1	EM	F1	EM	F1	EM		
CLaRa	16×	47.94	37.84	45.85	35.01	17.68	8.23	44.33	40.09	38.95	38.85
BIRC	16×	**55.19**	**44.83**	**52.71**	**41.29**	21.30	11.48	**48.80**	44.52	44.50	41.30
First-pass direct context	Text	54.31	44.16	51.87	41.07	**27.37**	**22.76**	48.05	**44.81**	**45.40**	**42.40**

A.27 MuSiQue Training-Partition Construction Control

BIRC's Phase II training uses the derived MuSiQue partition (168,745 examples; Appendix A.17), whereas the standard answerable partition contains 19,938 examples. To assess sensitivity to this choice, we retrain BIRC and CLaRa on the standard MuSiQue training partition (19,938 examples), while holding fixed the hyperparameters in Table A72, all Phase-I sources, and the NQ, HotpotQA, and 2WikiMultiHopQA Phase-II training splits.

With the standard partition, MuSiQue F1 is 19.24% for BIRC and 15.96% for CLaRa, giving a paired-seed BIRC–CLaRa difference of 3.28 pp. Under the derived partition, the corresponding difference is 3.62 pp (Table 3). For NQ, HotpotQA, and 2Wiki, every within-model derived-minus-standard F1 difference has absolute magnitude at most 0.34 pp. The BIRC–CLaRa Avg F1 differences are +5.57 pp with the standard partition and +5.55 pp with the derived partition.

Table A42: MuSiQue partition control (F1%, Normal, 16×; five-seed means). Standard/Derived: 19,938/168,745 training examples. Δ: Derived minus Standard. Bold marks within-model maxima.

Model	MuSiQue train	NQ	HotpotQA	MuSiQue	2Wiki	Avg F1
CLaRa	Derived (168k)	**47.94**	**45.85**	**17.68**	**44.33**	**38.95**
CLaRa	Standard (20k)	47.64	45.51	15.96	44.09	38.30
CLaRa	Δ	+0.30	+0.34	+1.72	+0.24	+0.65
BIRC	Derived (168k)	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**
BIRC	Standard (20k)	55.02	52.53	19.24	48.69	43.87
BIRC	Δ	+0.17	+0.18	+2.06	+0.11	+0.63

A.28 Hyperparameter Sensitivity Across Compression Ratios

Reported operating points. The primary K_{igfr} settings were chosen from benchmark evaluation results, with a separate choice at 128×. Tables A43 and A44 extend the 16× evaluation grids to 4× and 128×. The word "primary" identifies the reported configuration; it does not imply that the setting was fixed before observing these evaluation outcomes.

IGFR iteration count K_{igfr}. At 4×, $K_{\text{igfr}}=2$ has the highest Avg CEM in the grid. The difference between $K_{\text{igfr}}=2$ and $K_{\text{igfr}}=1$ is +0.33 pp, compared with +0.43 pp at 16× (41.30 vs. 40.87; Table A15). At 128×, $K_{\text{igfr}}=1$ is 0.16 pp higher in Avg CEM than $K_{\text{igfr}}=2$.

Table A43: IGFR sensitivity (CEM%, Normal, 4×/16×/128×); 16× repeats Table A15. Bold marks metric maxima and the primary iteration count.

CR	K_{igfr}	NQ	HotpotQA	MuSiQue	2Wiki	Avg
4×	1	59.49	46.98	11.88	48.93	41.82
	2	**59.86**	**47.35**	**12.19**	**49.20**	**42.15**
	3	59.58	47.16	11.96	48.82	41.88
16×	1	58.20	45.79	11.31	48.18	40.87
	2	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**
	3	58.55	46.02	11.49	48.22	41.07
128×	**1**	**54.12**	**41.72**	**9.63**	**42.33**	**36.95**
	2	53.84	41.56	9.47	42.29	36.79
	3	53.58	41.31	9.37	42.02	36.57

CFRS blending weight across compression ratios. Table A44 extends the 16× sweep from Table A16. We use the primary global setting $w_{\text{cfrs}} = 0.30$ at every compression ratio. The largest Avg CEM in the evaluation grid occurs at 0.20 for 4× and 128×; the default and 0.20 settings differ by 0.02 and 0.05 points, respectively.

Table A44: CFRS weight sensitivity (CEM%, Normal; five-seed means). $K_{\text{igfr}} = 2$ at 4×/16× and 1 at 128×. Bold marks within-CR Avg CEM maxima; † marks the primary weight.

CR	w_{cfrs}	NQ	HotpotQA	MuSiQue	2Wiki	Avg
4×	0.00	59.00	46.71	11.94	48.67	41.58
	0.10	59.42	47.06	12.02	49.30	41.95
	0.20	59.88	47.37	12.18	49.25	**42.17**
	$0.30^{\dagger}$	59.86	47.35	12.19	49.20	42.15
	0.40	59.54	47.08	11.96	49.14	41.93
	0.50	59.18	46.72	11.82	48.76	41.62
16×	0.00	57.48	45.42	11.29	48.01	40.55
	0.10	58.15	45.75	11.35	48.47	40.93
	0.20	58.60	46.00	11.55	48.49	41.16
	$0.30^{\dagger}$	58.73	46.17	11.74	48.56	**41.30**
	0.40	58.68	45.73	11.27	48.40	41.02
	0.50	57.82	45.28	11.08	48.50	40.67
128×	0.00	53.53	41.36	9.48	42.03	36.60
	0.10	53.92	41.62	9.55	42.31	36.85
	0.20	54.18	41.78	9.62	42.42	**37.00**
	$0.30^{\dagger}$	54.12	41.72	9.63	42.33	36.95
	0.40	53.80	41.45	9.42	42.13	36.70
	0.50	53.42	41.05	9.18	41.75	36.35

Across the three displayed ratios, the highest measured Avg CEM occurs at either $w_{\text{cfrs}} = 0.20$ or 0.30. Relative to no CFRS, the highest-scoring grid rows have Avg CEM values 0.59, 0.75, and 0.40 points higher at 4×, 16×, and 128×, respectively.

Cross-CR primary configuration. The primary setting is $K_{\text{igfr}} = 2$ at 4×–64× and $K_{\text{igfr}} = 1$ at 128×. At each displayed ratio, this selection picks the row with highest measured Avg CEM. All primary 128× BIRC CEM results (Table A2) use $K_{\text{igfr}} = 1$: NQ 54.12, HotpotQA 41.72, MuSiQue 9.63, 2Wiki 42.33, and Avg 36.95%. Its evaluation-set Avg CEM is 0.16 pp higher than that of $K_{\text{igfr}} = 2$. BIRC's Avg CEM of 36.95% exceeds CLaRa's 36.75% by 0.20 pp; the existing common-$K_{\text{igfr}} = 2$ row instead gives 36.79%, a 0.04-point difference. Both retain the evaluation-selected CFRS weight, so neither is an independent test of a development-only selection procedure. The interface-update ablation at 128× (Appendix A.19, Table A24) uses the same $K_{\text{igfr}} = 1$ as the main table.

A.29 CLaRa Mistral-7B: Five-Seed Baseline

For the training-schedule-matched baseline, we retrain CLaRa on Mistral-7B with five seeds at each of five compression ratios (25 runs). Training uses the same Phase-II question–answer examples, five-epoch schedule, and seeds as BIRC ($\{42, 123, 456, 789, 2024\}$), together with a learning rate of

2×10^{-4}, a 500-step warmup, a batch size of 32, and LoRA rank $r_{\mathrm{LoRA}}=16$. In the comparison, each method retains its native retrieval, objective, adapter placement, and initialization. All main-text CLaRa comparisons use five-seed means. For each ratio, the query-reasoner and reader adapters are initialized from the corresponding Phase-I compressor checkpoint, while the compressor and offline document representations remain fixed. Phase II minimizes answer next-token cross-entropy and updates the query reasoner and reader. CLaRa's straight-through selector uses hard top-k selection in the forward pass and soft-selection weights in the backward pass. Per-document memory counts depend only on truncated passage length and the shared nominal ratio, without score-conditioned allocation. This implements CLaRa's uniform-rate compression and joint selector–reader formulation (He et al., 2025) in the common passage format used for this study. Phase II presents five documents per query. Under Normal retrieval, the selector receives the frozen BGE retriever's top-20 list and passes its hard top-five selection to the reader. Under Oracle retrieval, CLaRa instead receives the same page-unique 100-page pool as BIRC and selects its hard top five from that pool. Recall uses the complete supporting-page annotations from the prepared benchmark split for both systems; supporting pages outside a system's selected set count as misses. Tables A45 and A46 report CEM% and F1%.

Table A45: CLaRa Mistral-7B CEM% (Normal/Oracle; five-seed means) under this study's data and training protocol.

Retr.	CR	NQ	HotpotQA	MuSiQue	2Wiki	Avg
Normal	4×	57.18	45.24	10.40	47.08	39.98
	16×	55.77	43.87	10.25	45.52	38.85
	32×	54.79	42.55	9.72	44.53	37.90
	64×	53.91	42.72	9.88	44.10	37.65
	128×	53.06	41.14	9.61	43.19	36.75
Oracle	4×	76.38	73.96	46.14	70.64	66.78
	16×	75.37	70.93	43.06	66.31	63.92
	32×	73.06	68.35	38.54	64.17	61.03
	64×	71.42	64.25	35.12	60.89	57.92
	128×	68.84	61.93	31.12	58.59	55.12

Table A46: CLaRa Mistral-7B F1% across compression ratios (Normal/Oracle; five-seed means).

Retr.	CR	NQ	HotpotQA	MuSiQue	2Wiki	Avg
Normal	4×	48.34	46.07	17.55	47.32	39.82
	16×	47.94	45.85	17.68	44.33	38.95
	32×	47.28	44.28	16.72	42.82	37.77
	64×	46.41	44.65	16.08	40.57	36.93
	128×	45.54	42.29	14.50	38.31	35.16
Oracle	4×	75.78	69.81	33.28	74.02	63.22
	16×	71.68	68.57	30.86	67.98	59.77
	32×	69.88	66.08	28.94	64.52	57.36
	64×	67.57	63.61	27.16	61.61	54.99
	128×	65.00	61.36	23.65	58.43	52.11

CLaRa EM results. Table A47 reports five-seed EM% under Normal retrieval. Across ratios, the five-seed SD is at most 0.58 pp for NQ EM, 1.14 pp for HotpotQA EM, 1.73 pp for MuSiQue EM, and 0.82 pp for 2Wiki EM. All main-text CLaRa rows use these five-seed means.

Table A47: CLaRa Mistral-7B EM% (Normal; five-seed means) under this study's data and training protocol.

Retr.	CR	NQ	HotpotQA	MuSiQue	2Wiki	Avg
Normal	4×	38.16	35.12	8.11	43.11	31.12
	16×	37.84	35.01	8.23	40.09	30.29
	32×	37.22	33.95	7.85	38.52	29.38
	64×	36.51	34.18	7.52	36.34	28.64
	128×	35.63	31.92	6.68	34.11	27.08

A.30 Per-Seed Intervals at 16×

Tables A48 and A49 report seed-level outcomes for the primary Full BIRC and CLaRa native-pipeline runs. The independently retrained interface-update suite in Appendix A.20 uses a separate set of independently trained Full checkpoints.

Table A48: Full BIRC per-seed results (16×, Normal). Scores and CI endpoints are in percent; SD is in percentage points. CIs use Student's t (4 df).

Dataset	Metric	42	123	456	789	2024	Mean	SD	95% CI (t, 4 df)
NQ	EM	44.60	45.03	44.83	45.17	44.50	44.83	0.28	[44.48, 45.18]
HotpotQA	EM	41.18	41.54	40.99	41.42	41.33	41.29	0.21	[41.03, 41.56]
MuSiQue	EM	11.20	11.74	11.53	11.28	11.65	11.48	0.23	[11.19, 11.77]
2Wiki	EM	44.39	44.64	44.29	44.79	44.48	44.52	0.20	[44.27, 44.77]
NQ	CEM	58.49	58.96	58.76	59.10	58.32	58.73	0.32	[58.33, 59.13]
HotpotQA	CEM	45.92	46.37	46.21	46.04	46.30	46.17	0.19	[45.94, 46.40]
MuSiQue	CEM	11.52	12.02	11.77	11.44	11.93	11.74	0.25	[11.42, 12.05]
2Wiki	CEM	48.42	48.71	48.51	48.83	48.34	48.56	0.20	[48.31, 48.81]
NQ	F1	55.45	55.01	54.88	55.50	55.11	55.19	0.27	[54.85, 55.53]
HotpotQA	F1	53.05	52.58	52.34	52.84	52.74	52.71	0.27	[52.38, 53.04]
MuSiQue	F1	21.34	21.37	21.06	21.01	21.72	21.30	0.28	[20.95, 21.65]
2Wiki	F1	49.20	48.52	48.36	49.13	48.79	48.80	0.37	[48.34, 49.26]
Avg	**EM**	35.34	35.74	35.41	35.67	35.49	**35.53**	**0.17**	[35.32, 35.74]
Avg	**CEM**	41.09	41.52	41.31	41.35	41.22	**41.30**	**0.16**	[41.10, 41.50]
Avg	**F1**	44.76	44.37	44.16	44.62	44.59	**44.50**	**0.24**	[44.21, 44.79]

Table A49: CLaRa per-seed results (16×, Normal). Scores and CI endpoints are in percent; SD is in percentage points. CIs use Student's t (4 df).

Dataset	Metric	42	123	456	789	2024	Mean	SD	95% CI (t, 4 df)
NQ	EM	37.62	38.03	37.79	38.17	37.57	37.84	0.26	[37.51, 38.16]
HotpotQA	EM	34.78	35.22	35.05	34.89	35.13	35.01	0.18	[34.79, 35.24]
MuSiQue	EM	7.94	8.48	8.32	7.86	8.56	8.23	0.32	[7.84, 8.62]
2Wiki	EM	39.98	40.24	40.07	40.35	39.82	40.09	0.21	[39.83, 40.35]
NQ	CEM	55.51	56.02	55.79	56.11	55.40	55.77	0.31	[55.38, 56.15]
HotpotQA	CEM	43.62	44.11	43.88	43.72	44.03	43.87	0.21	[43.62, 44.13]
MuSiQue	CEM	9.97	10.51	10.30	9.85	10.63	10.25	0.34	[9.83, 10.67]
2Wiki	CEM	45.39	45.68	45.48	45.81	45.26	45.52	0.22	[45.25, 45.80]
NQ	F1	48.14	47.78	47.65	48.54	47.59	47.94	0.40	[47.45, 48.43]
HotpotQA	F1	45.99	45.68	45.48	46.13	45.97	45.85	0.26	[45.52, 46.18]
MuSiQue	F1	17.60	17.74	17.35	17.61	18.10	17.68	0.27	[17.34, 18.02]
2Wiki	F1	44.67	44.16	44.32	44.60	43.90	44.33	0.32	[43.94, 44.72]
Avg	**EM**	30.08	30.49	30.31	30.32	30.27	**30.29**	**0.15**	[30.11, 30.48]
Avg	**CEM**	38.62	39.08	38.86	38.87	38.83	**38.85**	**0.16**	[38.65, 39.06]
Avg	**F1**	39.10	38.84	38.70	39.22	38.89	**38.95**	**0.21**	[38.69, 39.21]

For 2Wiki F1, the five seed-aligned BIRC–CLaRa differences are $4.53/4.36/4.04/4.53/4.89$ points (mean 4.47, sample SD 0.31); the two-sided 95% CI for the mean difference is $[4.09, 4.85]$.

A.31 SUPPORTING-BACKBONE INTERVENTION RESULTS ACROSS COMPRESSION RATIOS

Table A50 reports interventions on the five-stage supporting backbone at $16\times$; Tables A51 and A52 give the corresponding results at $4\times$ and $32\times$. Each BIRC control applies a test-time intervention to the same trained BIRC checkpoints, and downstream stages receive that intervention's outputs. Each table includes a separate native CLaRa reference. Across the three tables, bold cells mark within-column maxima. Here $-$QCA denotes complete component shutdown. It removes the route decision, QCA score $c(q)$, and subquestion outputs; because AHR no longer receives a query type, its type-specific weights are also inactive. In this component-shutdown suite, Full BIRC records 41.30 Avg CEM and $-$QCA records 40.20, a 1.10-point difference. By contrast, Table A1 forces every QCA route to Simple after surface QCA while retaining $c(q)$, the matched-rule trace, entity and hop counts, and template-derived subquestions; the corresponding Avg CEM difference is 0.31 points. Among the reported fixed-checkpoint BIRC interventions, the $-$MADS row has the largest Full-minus-intervention Avg CEM difference at every tested ratio. The NQ difference between Full and $-$QCA is $+0.38$ pp at $4\times$, $+1.32$ pp at $16\times$, and $+1.31$ pp at $32\times$.

Table A50: Fixed-checkpoint backbone interventions (CEM/F1%, Normal, $16\times$; five-seed means). Q/A/I/M/C: QCA/AHR/IGFR/MADS/CCEF. CLaRa is a native reference with no component status. Removing QCA retains AHR but disables its type-conditioned weights.

Variant	Q	A	I	M	C	CEM%					F1%				
						NQ	Hotpot	MuSi.	2Wiki	Avg	NQ	Hotpot	MuSi.	2Wiki	Avg
Full	✓	✓	✓	✓	✓	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**
−CCEF	✓	✓	✓	✓	—	57.89	45.61	10.83	48.07	40.60	54.15	52.03	20.31	48.19	43.67
−MADS	✓	✓	✓	—	✓	57.17	44.48	10.72	47.63	40.00	53.43	50.73	20.14	47.70	43.00
−AHR	✓	—	✓	✓	✓	57.83	45.09	10.84	47.84	40.40	54.13	51.36	20.24	47.87	43.40
−QCA	—	✓	✓	✓	✓	57.41	45.07	10.53	47.79	40.20	53.67	51.37	19.91	47.85	43.20
CLaRa (native reference)			*N/A*			55.77	43.87	10.25	45.52	38.85	47.94	45.85	17.68	44.33	38.95

Table A51: Fixed-checkpoint backbone interventions (CEM%, Normal, $4\times$; five-seed means). Q/A/I/M/C: QCA/AHR/IGFR/MADS/CCEF. CLaRa is a native reference with no component status.

Variant	Q	A	I	M	C	NQ	HotpotQA	MuSiQue	2Wiki
Full	✓	✓	✓	✓	✓	**59.86**	**47.35**	**12.19**	**49.20**
−CCEF	✓	✓	✓	✓	—	59.18	47.14	11.42	48.91
−MADS	✓	✓	✓	—	✓	57.94	45.96	11.75	47.56
−AHR	✓	—	✓	✓	✓	59.48	46.89	11.64	48.45
−QCA	—	✓	✓	✓	✓	59.48	47.09	11.84	49.08
CLaRa (native reference)			*N/A*			57.18	45.24	10.40	47.08

Table A52: Fixed-checkpoint backbone interventions (CEM%, Normal, $32\times$; five-seed means). Q/A/I/M/C: QCA/AHR/IGFR/MADS/CCEF. CLaRa is a native reference with no component status; bold marks maxima.

Variant	Q	A	I	M	C	NQ	HotpotQA	MuSiQue	2Wiki
Full	✓	✓	✓	✓	✓	**56.88**	**44.09**	**10.02**	**45.61**
−CCEF	✓	✓	✓	✓	—	56.29	43.89	9.73	45.17
−MADS	✓	✓	✓	—	✓	54.61	42.51	9.57	43.64
−AHR	✓	—	✓	✓	✓	56.01	43.14	9.73	44.79
−QCA	—	✓	✓	✓	✓	55.57	43.36	9.34	44.60
CLaRa (native reference)			*N/A*			54.79	42.55	9.72	44.53

A.32 MADS Scorer Combination Study Under Normal Retrieval

Appendix A.14 evaluates MADS scorer combinations under Oracle retrieval at 32×; Table A53 reports the analogous study under Normal retrieval at 16×. The full three-scorer configuration (Lex+Sem+Ent) achieves the highest Avg F1. Relative to Lex+Sem, it is 1.33 pp higher on HotpotQA and 0.87 pp higher on MuSiQue, versus 0.35 pp on NQ and 0.33 pp on 2Wiki. The primary configuration uses equal weights. For the development-selected-weight row, we form all within-query pairs comprising one gold-support candidate and one negative candidate in the same 500-query held-out development partition. We search a 0.01-spaced grid of nonnegative lexical, semantic, and entity weights that sum to one and select the vector minimizing the mean pairwise logistic ranking loss. Objective values within an absolute tolerance of 10^{-15} are treated as ties and resolved by selecting the lexicographically first vector. The selected vector is then fixed for test evaluation. As in the Oracle study, every row is a test-time scoring intervention on the same trained checkpoints.

Table A53: Fixed-checkpoint MADS combinations (F1%, Normal, 16×; five-seed means). Bold marks metric maxima and the primary equal-weight configuration.

Config	Lex	Sem	Ent	NQ	HotpotQA	MuSiQue	2Wiki	Avg F1
Lex only	✓	×	×	52.47	49.83	19.24	47.78	42.33
Sem only	×	✓	×	53.38	50.47	19.71	48.56	43.03
Lex+Sem	✓	✓	×	54.84	51.38	20.43	48.47	43.78
Lex+Sem+Ent	✓	✓	✓	55.19	**52.71**	**21.30**	48.80	**44.50**
Development-selected weights	✓	✓	✓	**55.47**	52.34	20.84	**48.83**	44.37

A.33 ACR Floor Sensitivity at 128×

Table A54 evaluates the ACR scale floor at 128× under Normal retrieval. All rows reuse the same five trained checkpoints and the main $K_{\text{igfr}} = 1$ configuration; only the evaluation-time scale floor changes. The route-adaptive rows keep $\rho^{\text{S}}_{\min} = 0.25$ for QCA-routed Simple queries and vary the Multi-Hop/Multi-Aspect floor over $\{0.35, 0.45, 0.55\}$. All remaining hyperparameters match Table A72.

The three route-adaptive settings yield Avg CEM values of 37.50–37.60%. Within this grid, $\rho^{\text{MH}}_{\min} = 0.45$ gives the highest Avg CEM (37.60%) and 2Wiki F1 (39.08%). The neighboring 0.35 and 0.55 settings give Avg CEM values of 37.50% and 37.55%, respectively. At this ratio, the mean soft cutoffs of later-ranked passages approach one state (Figure A4). Increasing the scale floor raises the corresponding gate weights while leaving the materialized memory lengths unchanged. These experiments measure test-time gate changes on the primary checkpoints; the table reports both aggregate and per-benchmark outcomes.

Table A54: ACR floor sensitivity (128×, $K_{\text{igfr}} = 1$, Normal; five-seed CEM%, with 2Wiki F1% last). S/MH denote Simple/shared Multi-Hop–Multi-Aspect floors; Uniform-scale instead fixes $\rho = 0.625$. Bold marks metric maxima.

Variant	$\rho^{\text{S}}_{\min}$	$\rho^{\text{MH}}_{\min}$	NQ	HotpotQA	MuSiQue	2Wiki	Avg CEM	2Wiki F1
Full BIRC	0.25	0.25	**54.12**	41.72	9.63	42.33	36.95	38.02
Uniform-scale control	0.625	0.625	52.89	42.17	10.08	44.26	37.35	38.37
Adaptive	0.25	0.35	54.03	42.18	9.93	43.86	37.50	38.62
Adaptive	0.25	0.45	53.92	**42.32**	**10.11**	44.05	**37.60**	**39.08**
Adaptive	0.25	0.55	53.67	42.22	10.03	**44.28**	37.55	38.93

A.34 Cross-Reader Results: Qwen2.5-14B-Instruct

We evaluate Qwen2.5-14B-Instruct (Qwen et al., 2024) at 16× under Normal retrieval. Training follows the two-phase protocol used in the main experiments—including the same derived MuSiQue training partition (Appendix A.17)—with these adjustments for the larger backbone: an effective batch

size of 16 with gradient accumulation over two steps, a $5120 \to 2560 \to 1024$ MTFRL projection, and a learning rate of 1.6×10^{-4} (0.8 times the Mistral/Llama rate). LoRA rank remains 16, and the warmup length and epoch counts are unchanged. We train both BIRC and CLaRa on the augmented data using each of the five seeds.

BIRC achieves 42.30% Avg CEM ($+1.83$ pp relative to CLaRa) and 45.60% Avg F1 ($+4.49$ pp; Table A55). Across Mistral-7B, Llama-3-8B, and Qwen2.5-14B, the corresponding improvements span $+1.83$–$+2.92$ pp in Avg CEM and $+4.49$–$+6.16$ pp in Avg F1. The Qwen2.5-14B CLaRa baseline achieves 40.47% Avg CEM, compared with 38.85% for its Mistral-7B counterpart. Table A55 reports the five-seed endpoint means used in the cross-backbone comparison.

Table A55: Qwen2.5-14B-Instruct results (CEM/F1%, Normal, 16×; five-seed means). Δ: BIRC minus CLaRa (pp). Bold marks maxima.

Model	CEM%					F1%				
	NQ	Hotpot	MuSiQue	2Wiki	Avg	NQ	Hotpot	MuSiQue	2Wiki	Avg
CLaRa	57.87	45.48	11.38	47.16	40.47	50.15	47.99	19.76	46.54	41.11
BIRC	**59.79**	**47.32**	**12.61**	**49.48**	**42.30**	**55.87**	**53.36**	**22.68**	**50.49**	**45.60**
Δ	+1.92	+1.84	+1.23	+2.32	+1.83	+5.72	+5.37	+2.92	+3.95	+4.49

A.35 Frozen Llama-3-70B-Instruct Reader at 16×

We evaluate the frozen `meta-llama/Meta-Llama-3-70B-Instruct` reader (Grattafiori et al., 2024) at 16× under Normal retrieval on the same four benchmarks. The reader is fixed; each system contributes five seed-specific retrieval–compression checkpoints. This evaluation complements the three jointly trained readers in Table 4.

The reader is loaded with 4-bit NF4 quantization and bfloat16 computation, with NF4 double quantization disabled. Inputs are capped at 1,024 tokens within the native 8,192-token context window. Greedy decoding uses a repetition penalty of 1.0 and stops at the end-of-sequence or end-of-turn token (IDs 128001 and 128009), or after 64 new tokens. Within each seed, BIRC and CLaRa use the same frozen reader. The evaluation uses a batch size of 1 on a single `NVIDIA A100-SXM4-80GB` GPU with Python 3.10, CUDA 12.1, PyTorch 2.3.1, Transformers 4.43.3, and bitsandbytes 0.43.3.

Table A56: Frozen Llama-3-70B-Instruct reader (16×, Normal). (a) Five-seed means; bold marks task-score maxima. (b) Per-seed Avg F1 with 95% CIs (Student's t, 4 df).

(a) Benchmark endpoints

System	NQ F1	Hotpot F1	MuSiQue F1	2Wiki F1	Avg F1	Avg CEM	Sec./query
CLaRa	54.13	51.82	25.18	56.07	46.80	44.60	18.43
BIRC	**59.78**	**57.61**	**29.08**	**59.13**	**51.40**	**46.80**	19.27
Δ	+5.65	+5.79	+3.90	+3.06	+4.60	+2.20	+0.84

(b) Per-seed Avg F1

System	42	123	456	789	2024	Mean	SD	95% CI (t, 4 df)
CLaRa	47.02	46.63	46.48	46.95	46.92	46.80	0.23	[46.51,47.09]
BIRC	51.37	51.43	50.93	51.80	51.47	51.40	0.31	[51.01,51.79]

The paired Avg F1 difference is $+4.60$ percentage points (95% CI from the benchmark-stratified paired query-level bootstrap $[4.26, 4.94]$). The four benchmark differences are $+5.65$, $+5.79$, $+3.90$, and $+3.06$ percentage points.

A.36 Normal Retrieval Recall@5 Across Compression Ratios

Table A57: Native first-pass Recall@5% by nominal CR, with separately trained checkpoints. BIRC is measured before compression/MTFRL/CFRS; CLaRa selects five from frozen-BGE top 20 using compressed representations. Pools and selectors are system-specific (Appendix A.22). Avg and Δ use displayed two-decimal values; bold marks within-CR maxima.

Model	Nominal CR	NQ	HotpotQA	MuSiQue	2Wiki	Avg
CLaRa	4×	72.48	60.05	35.87	66.34	58.69
CLaRa	16×	71.18	58.41	34.15	64.73	57.12
CLaRa	32×	69.99	56.82	33.01	63.43	55.81
CLaRa	64×	68.46	55.16	31.44	61.93	54.25
CLaRa	128×	66.76	53.54	29.99	60.12	52.60
BIRC	4×	**75.42**	**65.36**	**39.22**	**70.76**	**62.69**
BIRC	16×	**74.20**	**64.10**	**38.10**	**68.40**	**61.20**
BIRC	32×	**72.98**	**62.86**	**36.92**	**66.68**	**59.86**
BIRC	64×	**71.62**	**61.42**	**35.48**	**64.48**	**58.25**
BIRC	128×	**70.13**	**59.88**	**33.94**	**62.25**	**56.55**
Δ (BIRC−CLaRa)	4×	+2.94	+5.31	+3.35	+4.42	+4.00
	16×	+3.02	+5.69	+3.95	+3.67	+4.08
	32×	+2.99	+6.04	+3.91	+3.25	+4.05
	64×	+3.16	+6.26	+4.04	+2.55	+4.00
	128×	+3.37	+6.34	+3.95	+2.13	+3.95

Table A57 summarizes first-pass performance across the five nominal ratios. The QR LoRA adapter is updated in Phase II, with a separate checkpoint trained for each ratio. Before MTFRL and CFRS, the average BIRC–CLaRa Recall@5 difference ranges from $+3.95$ to $+4.08$ percentage points.

A.37 First-Pass Retrieval Before and After Phase II

We evaluate the $16\times$ first-pass output of the same retrieval pipeline immediately before and after Phase II QA training. Both measurements follow AHR–IGFR–MADS–CCEF and precede compression, MTFRL, and CFRS. The macro-averaged post-minus-pre differences are 1.38, 1.53, and 1.50 percentage points at Recall@1, Recall@3, and Recall@5, respectively (Table A58).

Before Phase II, the query reasoner is initialized from the paired Phase-I compressor adapter. Both endpoints use the same frozen W_{BGE} and share the corpus, indices, BM25 and MADS configurations, and first-pass evaluation pipeline.

Table A58: First-pass Recall@k% before/after Phase II (nominal $16\times$), measured before compression/MTFRL/CFRS. Δ: Post minus Pre. Averages use full-precision values.

	Recall@1			Recall@3			Recall@5		
Dataset	**Pre**	**Post**	Δ	**Pre**	**Post**	Δ	**Pre**	**Post**	Δ
NQ	38.6	40.0	+1.4	63.8	65.4	+1.6	72.8	74.2	+1.4
HotpotQA	29.6	31.2	+1.6	52.7	54.5	+1.8	62.3	64.1	+1.8
MuSiQue	16.0	17.1	+1.1	29.2	30.3	+1.1	36.9	38.1	+1.2
2Wiki	32.2	33.6	+1.4	57.6	59.2	+1.6	66.8	68.4	+1.6
Avg	**29.10**	**30.48**	**+1.38**	**50.82**	**52.35**	**+1.53**	**59.70**	**61.20**	**+1.50**

Across five seeds, the per-benchmark Recall@5 standard deviations range from 0.5 to 0.9 percentage points before Phase II and from 0.5 to 0.8 after Phase II; all four paired query-level bootstrap intervals exclude zero (Table A59). The intervals are computed over paired evaluation items within each benchmark. This diagnostic reports population SD across its five paired checkpoints. Elsewhere, seed-level inference uses sample SD and Student-t intervals with four degrees of freedom.

Table A59: Recall@5 before/after Phase II (16×): five-seed mean ± population SD. Intervals use 10,000 paired query resamples per benchmark after averaging aligned seed-pair differences. Δ: Post minus Pre.

Dataset	Pre R@5	Post R@5	Δ	95% CI	p
NQ	72.8 ± 0.5	74.2 ± 0.5	+1.4	$[+0.8, +2.0]$	< 0.01
HotpotQA	62.3 ± 0.8	64.1 ± 0.7	+1.8	$[+1.1, +2.5]$	< 0.001
MuSiQue	36.9 ± 0.9	38.1 ± 0.8	+1.2	$[+0.3, +2.1]$	< 0.05
2Wiki	66.8 ± 0.6	68.4 ± 0.5	+1.6	$[+1.0, +2.2]$	< 0.001

A.38 QCA Sensitivity Analysis: Rule-Based vs. Zero-Shot LLM Routing

We compare BIRC's rule-based QCA with two protocols that use the same adapter-free, zero-shot Mistral-7B router. For each protocol, the router runs once per distinct applicable query, and its label is cached by benchmark and question ID. Both protocols cover all evaluation queries; train+test additionally covers Phase-II training queries. The 1,000-query reference panel is used only to compute routing weighted F1. The label space is {SIMPLE, MULTI-ASPECT, MULTI-HOP}, and routing uses neither label supervision nor in-context examples. Only the routed category changes; the surface-rule pass continues to provide $c(q)$, entity counts, rule traces, and deterministic subquestion templates. The versioned prompt defines the three labels and requires an exact first line, `Label: <simple|multi-aspect|multi-hop>`, followed by a nonempty `Rationale:` line. Decoding is greedy to the end-of-sequence token or 64 new tokens with all adapters disabled. The protocol validates each response against the prescribed label set and two-line format.

The *test-only* protocol substitutes cached LLM routes into the five rule-trained checkpoints, leaving all downstream parameters fixed. The *train+test* protocol independently trains five downstream checkpoints with the adapter-free router used during both Phase II and evaluation. At evaluation, both protocols reuse the same evaluation-route cache across seeds; the five train+test runs also reuse the Phase-II training-route cache. QA metrics use the full four-benchmark evaluation at 16× under Normal retrieval. By contrast, AdaptiveRAG trains a query-complexity classifier on automatically constructed labels and uses different routing strategies and supervision (Jeong et al., 2024).

Table A60: Rule-based versus zero-shot LLM routing (Mistral-7B). QCA weighted F1: 1,000-query panel; QA: Normal, 16×, five-seed means. Latency: batch-1 wall-clock seconds/query. Δ Avg F1 is relative to rule-based train+test.

Protocol	QCA wF1	Router ms	Sec./query	Avg CEM	Avg F1	Δ F1
Rule-based train+test	84.12	0.73	6.62	41.30	44.50	0.00
LLM test-only	87.47	312.08	6.94	41.43	44.60	+0.10
LLM train+test	87.47	312.08	6.94	41.55	44.75	+0.25

Table A61: Router protocols: per-seed Avg F1 and 95% CIs (Student's t, 4 df).

Protocol	42	123	456	789	2024	Mean	SD	95% CI
Rule-based train+test	44.76	44.37	44.16	44.62	44.59	44.50	0.24	[44.21,44.79]
LLM test-only	44.74	44.59	44.20	44.80	44.67	44.60	0.24	[44.31,44.89]
LLM train+test	44.81	44.79	44.34	44.97	44.84	44.75	0.24	[44.45,45.05]

Results. Rule-based QCA records 84.12% weighted routing F1, 0.73 ms router latency, and 44.50% Avg F1. The LLM test-only and train+test protocols both record 87.47% weighted routing F1 and 312.08 ms router latency, with Avg F1 scores of 44.60% and 44.75%, respectively. Their Avg F1 differences from rule-based QCA are $+0.10$ and $+0.25$ pp, with an additional 0.32 seconds per query. Rule-based QCA provides the lower-cost routing option. The LLM controls have higher routing and QA F1, together with longer measured latency, within the same retrieval–compression pipeline.

A.39 FIXED-CHECKPOINT ALL-UPDATES-OFF ENDPOINT COMPARISON

The All-updates-off variant retains BIRC's five-stage retrieval pipeline but disables all three interface updates. It passes all $K_i^{(0)}$ base states unchanged instead of applying ACR, uses CCEF ordering without CFRS, and omits the MTFRL second round. At nominal $16\times$, both this variant and Full BIRC materialize and process a mean of 184 reader memory positions per query. For Full BIRC, the corresponding mean gate-induced prefix total is 108; this statistic summarizes the gate profile. Results are five-seed means from the Full-BIRC checkpoints with Mistral-7B at $16\times$ under Normal retrieval. This variant matches the "All updates off" configuration in Table 1. Table A62 places its endpoint alongside those of CLaRa and Full BIRC.

Table A62: CLaRa, fixed-checkpoint all-updates-off, and Full BIRC (CEM/F1%, Normal, $16\times$; five-seed means). Bold marks column maxima.

System	CEM%					F1%				
	NQ	Hotpot	MuSiQue	2Wiki	Avg	NQ	Hotpot	MuSiQue	2Wiki	Avg
CLaRa	55.77	43.87	10.25	45.52	38.85	47.94	45.85	17.68	44.33	38.95
BIRC (all updates off)	56.03	43.82	10.34	46.21	39.10	51.93	49.17	18.34	45.76	41.30
Full BIRC	**58.73**	**46.17**	**11.74**	**48.56**	**41.30**	**55.19**	**52.71**	**21.30**	**48.80**	**44.50**

A.40 GATE-INDUCED PREFIX STATISTICS AND PHYSICAL ALLOCATION

After each passage is truncated to 768 passage-content tokens, the $1{,}000$-query sample contains a rounded mean of $2{,}950$ non-structural source tokens per query, or 590 tokens per passage. CLaRa materializes $K_i^{(0)}(r) = \max\{1, \lfloor L_i/r \rfloor\}$ memory positions for each of the five retrieved documents, for a total of $\sum_i K_i^{(0)}(r)$ positions. The empirical mean totals are $738/184/92/46/23$ positions at nominal $4/16/32/64/128\times$.[1] Full BIRC materializes the same base-position totals, but ACR uses $\rho_i \in [\rho_{\min}, \rho_{\max}]$ to construct the token-wise gates $g_i^{(t)}$ in Eq. 4 and defines T_i as a gate-induced prefix count. Under the measured profile (Table 5), $\sum_i \rho_i = 2.93$ on average and the gate-induced prefix totals are $432/108/54/27/13$ at nominal $4/16/32/64/128\times$. The gate-induced prefix count is $108/184 = 58.7\%$ of the physical count; the reader processes all 184 base positions.

Per-document gate statistics. Table 5 reports the empirical gate-induced prefix distribution at nominal $16\times$, averaged over $1{,}000$ queries (250 per benchmark). All queries in this diagnostic sample retained five CCEF survivors, yielding $N = 5{,}000$ document slots. Displayed ranks follow the post-CFRS reader order. Each ρ_i is computed from the fused retrieval score of the corresponding final candidate before CFRS and remains attached to that document during CFRS sorting. The rounded mean is $\rho_1 = 0.98$ for the first-ranked document and $\rho_5 = 0.31$ for the fifth (Tables 5 and A63). Dividing the rounded global mean source length $L = 590$ by each rank's mean T_i gives the displayed ratios, from $16.3\times$ at rank 1 to $51.8\times$ at rank 5 (harmonic mean $27.3\times$, arithmetic mean $32.6\times$). These diagnostic ratios summarize the gate profile. ACR attenuates the memory states, and the Full BIRC reader receives all base positions, averaging 184 per query. Figure A4 reports the measured per-rank soft-cutoff means from the corresponding evaluation runs at nominal ratios from $16\times$ to $128\times$. Table A63 provides the per-rank standard deviations and aggregate values underlying the compact main-text panel.

[1] The empirical totals average $\sum_i K_i^{(0)}(r)$ over the per-document context-length distribution, with flooring and the one-state lower bound applied before averaging. The per-rank T_i values in Table 5 are averaged in the same way.

Table A63: Gate-induced prefix statistics (16×; 1,000 queries). Post-CFRS ranks retain each document's retrieval-derived scale ρ_i; T_i is diagnostic, with all states materialized. Ratios divide rounded mean $L = 590$ by rank-mean T_i. Aggregate gives mean ρ_i, mean T_i, and the harmonic-mean ratio.

Final rank	**Mean ρ_i**	**SD ρ_i**	**Mean T_i**	**Diagnostic ratio**
1	0.98	0.04	36.1	16.3×
2	0.74	0.09	27.2	21.7×
3	0.52	0.11	19.1	30.9×
4	0.38	0.08	14.0	42.1×
5	0.31	0.06	11.4	51.8×
Aggregate	*0.586*	—	*21.6*	*27.3× harm.*

A.40.1 Matched Physical-Position Control

This control compares physical ACR with uniform allocation at the same per-query sequence length, averaging 108.0 positions. Both arms use BIRC's five first-pass CCEF survivors and their fused scores, with MTFRL disabled. This pool differs from the frozen-BGE top-five set used for the CLaRa comparison (Appendix A.23). Within each seed, both arms share a checkpoint trained with physical ACR and differ only in evaluation-time allocation. Tables A64 and A65 report five-seed results and intervals.

Table A64: Physical ACR versus per-query length-matched uniform allocation (16×, Normal). Both share checkpoints, first-pass passages/scores, and the MTFRL-off path, averaging 108.0 positions. Task metrics: %; Δ: ACR minus uniform. Only task metrics are bold.

Configuration	**NQ F1**	**Hotpot F1**	**MuSiQue F1**	**2Wiki F1**	**Avg EM**	**Avg CEM**	**Avg F1**	**Positions**	**Sec./query**	**Peak GB**
Physical ACR	**55.09**	**52.58**	**21.17**	**48.76**	**35.40**	**41.15**	**44.40**	108.0	6.15	20.6
Matched physical uniform	54.57	51.76	20.23	47.84	34.70	40.45	43.60	108.0	6.14	20.6
Δ ACR−uniform	**+0.52**	**+0.82**	**+0.94**	**+0.92**	**+0.70**	**+0.70**	**+0.80**	0.0	+0.01	0.0

Table A65: Physical-allocation controls: per-seed Avg F1, sample SD, and 95% CIs (Student's t, 4 df). Each pair shares a checkpoint; the final row is ACR minus uniform.

Configuration	**42**	**123**	**456**	**789**	**2024**	**Mean**	**SD**	**95% CI (t, 4 df)**
Physical ACR	44.15	44.48	44.76	44.05	44.56	44.40	0.29	$[44.03, 44.77]$
Matched uniform	43.50	43.53	44.04	43.17	43.76	43.60	0.32	$[43.20, 44.00]$
ACR−uniform	+0.65	+0.95	+0.72	+0.88	+0.80	+0.80	0.12	$[+0.65, +0.95]$

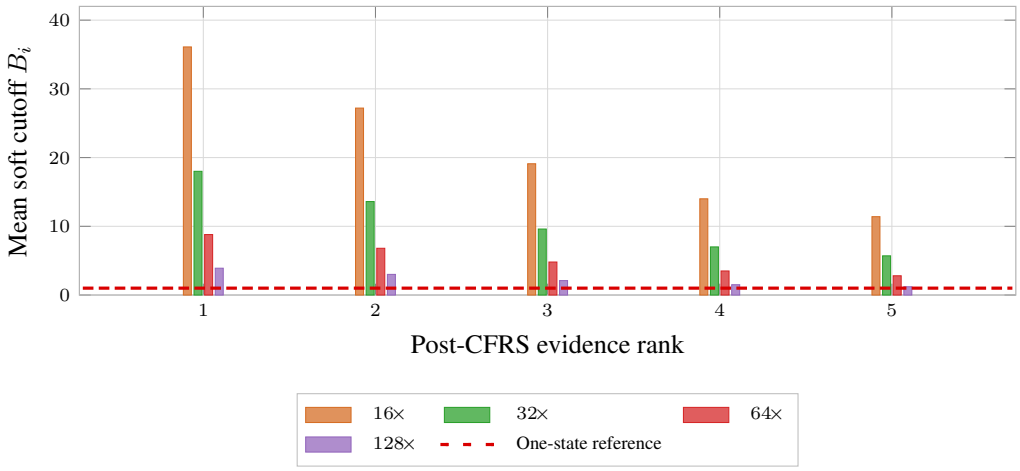

Figure A4: Mean ACR soft cutoff $B_i = \rho_i K_i^{(0)}(r)$ by post-CFRS rank at 16×–128× (1,000 queries per ratio). The dashed line marks one state. Soft cutoffs precede per-query flooring and the one-state minimum.

A.41 Candidate-Budget Grid

Table A66 applies test-time changes to the AHR candidate-pool and MTFRL feedback-pool budgets of Full BIRC checkpoints at 16× under Normal retrieval, with the IGFR configuration fixed. All rows use the complete four-benchmark panel at batch size 1. The $4{,}000/200$ row is the primary Full BIRC operating point.

Table A66: AHR/MTFRL budget grid (16×, Normal; IGFR fixed). Task scores: five-seed means with equal benchmark weights; Final R@5: recall on the final five passages (%). Latency: query-weighted seconds/query across 20 seed–benchmark runs; peak memory: maximum GB across runs. Δ Avg F1 is relative to $4{,}000/200$ (pp), the bold primary budget.

AHR pool	MTFRL pool	Avg CEM	Avg F1	Final R@5 (%)	Sec./query	Peak GB	Δ Avg F1 (pp)
20	0	39.20	42.20	55.3	5.92	20.8	−2.30
20	50	39.55	42.55	57.0	6.02	20.9	−1.95
20	200	39.88	42.82	58.1	6.15	20.9	−1.68
100	0	39.70	42.68	57.5	5.98	20.9	−1.82
100	50	40.05	43.08	59.3	6.08	21.0	−1.42
100	200	40.30	43.32	60.3	6.21	21.0	−1.18
1000	0	40.30	43.38	60.1	6.12	21.0	−1.12
1000	50	40.70	43.78	62.1	6.24	21.1	−0.72
1000	200	41.02	44.12	63.1	6.37	21.1	−0.38
4000	0	40.42	43.62	61.2	6.36	21.1	−0.88
4000	50	41.02	44.22	63.2	6.47	21.2	−0.28
4000	**200**	41.30	44.50	64.5	6.62	21.2	+0.00

A.42 Execution Profiles and Measurement Boundaries

Table A67 gives mean per-query latency and peak allocated memory at batch size 1 over 1,000 queries on a single NVIDIA A100 80 GB GPU. At 16×, BIRC's native profile records 6.62 seconds per query, including online candidate acquisition. CLaRa's native profile records 5.83 seconds per query; its timing begins with materialized candidates. For BIRC, the routing and scoring categories in the table take 0.38 and 0.24 seconds per query, respectively, including the feedback round. With the AHR pool fixed at 4,000, adding the 200-candidate feedback pool increases Avg F1 by 0.88 percentage points and latency by 0.26 seconds (6.36 to 6.62) relative to the $4{,}000/0$ setting. The matched physical control runs at 6.15 seconds under a different intervention: it shortens the memory sequence and disables MTFRL.

Timing begins after the checkpoints are loaded and the corpus index or candidate set is prepared. The frozen BGE direct-context row likewise begins from materialized candidates. The within-BIRC candidate-budget and physical-length controls provide matched latency contrasts. The main text additionally reports 6.00 seconds for BIRC's compression/CFRS plus decoder subtotal and 5.83 seconds for CLaRa's selection/compression plus decoder subtotal, computed from this table. These are diagnostic stage sums, not newly timed runs from identical cached candidates. Query encoding is charged to Routing for BIRC and Compression for CLaRa, and the candidate sources and encoded workloads differ. Subtracting BIRC's routing/scoring time therefore does not produce a fully matched compression benchmark. Peak memory is measured over the complete recorded invocation and cannot be adjusted by subtracting stage times. The candidate-budget grid in Table A66 uses the complete evaluation panels and its stated query-weighted aggregation.

Table A67: Batch-1 latency over 1,000 queries, peak allocated CUDA memory (decimal GB), and evidence positions. Frozen-BGE direct context reads BGE top five, excludes retrieval timing, and differs from first-pass CCEF direct context. Routing: query encoding/QCA/AHR/IGFR/MTFRL; Scoring: MADS/CCEF and union rescoring; Comp.: compression/CFRS/CLaRa selection; Dec.: input preparation/decoding. Dashes indicate stages not included as separate timing categories. Positions count source or latent tokens, excluding query/instruction; $\sum_i T_i$ is BIRC's diagnostic prefix total. Latency and position counts are query means.

System	CR	Comp.	Routing	Scoring	Dec.	Recorded total	Peak GB	Physical evidence positions	Gate-induced prefix count
Frozen-BGE direct context	1×	—	—	—	8.43	8.43	32.4	2,950	—
CLaRa	4×	3.21	—	—	3.32	6.53	23.7	738	—
CLaRa	16×	2.54	—	—	3.29	5.83	20.8	184	—
CLaRa	32×	2.17	—	—	3.28	5.45	19.9	92	—
CLaRa	64×	1.89	—	—	3.28	5.17	19.3	46	—
CLaRa	128×	1.48	—	—	3.28	4.76	18.9	23	—
BIRC	4×	3.05	0.42	0.27	3.67	7.41	24.1	738	432
BIRC	16×	2.68	0.38	0.24	3.32	6.62	21.2	184	108
BIRC	32×	2.31	0.35	0.22	3.30	6.18	20.3	92	54
BIRC	64×	1.98	0.33	0.21	3.29	5.81	19.6	46	27
BIRC	128×	1.63	0.31	0.19	3.28	5.41	19.2	23	13

A.43 MuSiQue Derived Training Partition

Phase II consumes a versioned $168{,}745$-row derived MuSiQue source comprising $19{,}938$ original, $52{,}107$ subquestion, $70{,}845$ partial-chain, and $25{,}855$ entity-variant rows. Before training, we verified family membership, unique row IDs, parent links, preserved decompositions, parent answers for partial-chain and entity variants, and annotated-hop answers for subquestion rows. Entity variants also passed checks for exact labeled-entity-surface preservation under spaCy and ROUGE-L ≥ 0.4 against the parent question.

A.44 Evaluation-Metric Definitions and Normalization

BIRC, our five-seed CLaRa baseline, and the complete SARA, SeleCom, OSCAR, and COCOM output panels in Table A5 use a common CEM normalization before substring matching. In Tables A2 and A4, the remaining literature-reference rows retain the values and normalization conventions from CLaRa's comparison table. For a model output $\hat{a}$ and benchmark reference a^*, we apply the following operations to both strings: (1) Unicode NFKC normalization; (2) lowercasing; (3) removing the English articles `a`, `an`, and `the` when they occur as standalone tokens; (4) stripping ASCII punctuation except apostrophes and hyphens; and (5) collapsing whitespace to single spaces. CEM is then defined as $\mathbf{1}[\mathrm{Substr}(a^*_{\mathrm{norm}}, \hat{a}_{\mathrm{norm}})]$, where $\mathrm{Substr}(x, y)$ means that x occurs contiguously in y; EM uses equality instead. For an empty normalized reference, CEM is one only when the normalized output is also empty. Token-level F1 applies the same normalization to multisets of whitespace-delimited tokens. F1 equals one when both output and reference are empty and zero when only one is empty. Each prepared evaluation record stores a single reference answer used for every system and seed. For retrieval, let $R_q = (d_1, \ldots, d_5)$ denote the system-returned list for query q. Let $\widetilde{R}_q$ retain the first occurrence of each page ID within R_q; duplicates are removed without filling from lower ranks. Then $D_q^{(k)}$ is the first $\min(k, |\widetilde{R}_q|)$ items of $\widetilde{R}_q$, and $P_q = \mathrm{pid}(\mathcal{P}(q))$, where $\mathcal{P}(q)$ is the set of annotated

supporting passages and pid maps passages to their distinct page IDs. We report supporting-page recall as

$$\text{Recall@}\,k = \frac{1}{|\mathcal{Q}_{\text{sup}}|} \sum_{q \in \mathcal{Q}_{\text{sup}}} \frac{|\,\text{pid}(D_q^{(k)}) \cap P_q|}{|P_q|}, \tag{24}$$

where $\mathcal{Q}_{\text{sup}}$ contains evaluation queries with at least one annotated supporting page. Benchmark-level Recall@k is this query mean, and the cross-benchmark average is unweighted. All tabulated Recall@k values are the resulting proportions multiplied by 100. The canonical reference answer is fixed before scoring. The alias-sensitivity evaluator joins each alias set to the evaluation panel by example ID and maximizes each metric over the fixed reference and its aligned aliases. Token-level F1 is the primary metric; CEM is a secondary normalized-reference substring metric.

A.45 Alias-Aware Scoring Sensitivity for Full BIRC

Table A68 reports the same Full BIRC outputs scored either against the single reference or against that reference and its aliases, taking the maximum for each metric across valid answers. Avg EM, Avg CEM, and Avg F1 increase by 0.88, 0.78, and 0.43 pp. This analysis measures the sensitivity of absolute scores to answer aliases; cross-system comparisons use the single-reference metric.

Table A68: Full BIRC alias scoring (16×, Normal). Δ Avg: alias-aware minus single-reference (pp). Bold marks panel maxima.

(a) EM

Scoring	NQ	HotpotQA	MuSiQue	2Wiki	Avg	ΔAvg
Single reference	44.83	41.29	11.48	44.52	35.53	—
Alias-aware	**46.17**	**41.91**	**11.99**	**45.55**	**36.41**	**+0.88**

(b) CEM

Scoring	NQ	HotpotQA	MuSiQue	2Wiki	Avg	ΔAvg
Single reference	58.73	46.17	11.74	48.56	41.30	—
Alias-aware	**59.80**	**46.78**	**12.31**	**49.42**	**42.08**	**+0.78**

(c) Token F1

Scoring	NQ	HotpotQA	MuSiQue	2Wiki	Avg	ΔAvg
Single reference	55.19	52.71	21.30	48.80	44.50	—
Alias-aware	**55.70**	**53.07**	**21.72**	**49.22**	**44.93**	**+0.43**

A.46 Train–Test Overlap Audit and Exclusion

We exclude Phase-I and Phase-II training rows whose canonical page URLs occur in the evaluation set. URL canonicalization lowercases schemes and hosts, collapses repeated path separators, and removes trailing slashes and fragments. We verify the post-exclusion source counts before constructing the training sets. Phase-II retrieval uses a separate training corpus with the same test-page exclusion.

The Phase-II audit compares all 485,006 remaining training items with supporting-passage texts from every evaluation split. The exact-page screen also checks stripped page IDs. Normalized 13-grams use an NFKC-normalized, case-folded alphanumeric token stream. Entity matching flags an item when an unordered pair of distinct normalized entity surfaces co-occurs within one sentence in that training passage and within one sentence in an evaluation passage; the retained spaCy types are PERSON, ORG, GPE, LOC, FAC, EVENT, PRODUCT, and WORK_OF_ART. The semantic screen assigns each training item its maximum raw English `roberta-large` BERTScore F1 over all evaluation passages, without IDF weighting or baseline rescaling. The exhaustive comparison uses deterministic shards without a retrieval or lexical prefilter.

Candidate rates are the fractions of training items flagged by each screen. We reviewed every item flagged by the 13-gram, entity-tuple, or BERTScore-≥ 0.90 screen and recorded one decision per item. The adjudication criterion is whether an overlap contains an evaluation answer or reasoning

chain; topical or entity overlap alone does not qualify. Lexical and entity review uses all matching evaluation passages; semantic review uses the maximum-scoring passage, with ties broken by evaluation ID. Candidate and adjudicated overlap rates both use the $485{,}006$-item denominator. The separate BERTScore-≥ 0.80 diagnostic flags 2.03% of items. Dedup+ removes the union of items adjudicated as substantive overlaps.

After URL exclusion, the exact-page screen finds no residual canonical-URL or stripped-page-ID matches. For the reviewed screens, candidate rates range from 0.14% to 0.76% and adjudicated overlap rates from 0.02% to 0.05% (Table A69).

Table A69: Phase-II overlap after exact page exclusion: all 485,006 training items form the denominator. Semantic review uses a BERTScore F1 ≥ 0.90 screen.

Check	**Candidate rate (%)**	**Adjudicated overlap (%)**
Exact page URL / page ID after exclusion	0.00	0.00
Normalized 13-gram	0.14	0.02
Entity tuple	0.76	0.03
BERTScore F1 ≥ 0.90	0.32	0.05

For Phase I, we deterministically sample $50{,}000$ rows from each of the four training categories and verify membership before scoring. The audit uses the same maximum-BERTScore rule, 0.90 threshold, and item-level review as Phase II. The aggregate candidate rate is 0.20%, and manual review identifies 0.04% of sampled rows as substantive overlaps (Table A70). This $200{,}000$-row semantic audit covers approximately 2.56% of the $7{,}808{,}465$ Phase-I pairs, sampled equally by category rather than in proportion to category size. Its aggregate overlap rate is a category-balanced sample statistic, not a full-corpus prevalence estimate. Exact URL exclusion is exhaustive; a full Phase-I 13-gram screen and semantic review are not established by this sample. Dedup+ below consequently does not certify that all Phase-I overlap has been removed.

Table A70: Phase-I semantic overlap (% within each 50,000-item category sample).

Category	n	**BERTScore** ≥ 0.90 **(%)**	**Adjudicated overlap (%)**
SimpleQA	50,000	0.18	0.03
ComplexQA	50,000	0.25	0.05
Paraphrase	50,000	0.14	0.03
Entity-Augmented	50,000	0.22	0.05
Overall	**200,000**	**0.20**	**0.04**

Dedup+ removes the IDs of confirmed overlaps from the exhaustive Phase-II screens and the 200,000-row Phase-I sample, then follows the primary $16\times$ Normal-retrieval training protocol. Its Avg CEM and Avg F1 are 0.06 and 0.02 percentage points below the full-training-set values, respectively (Table A71).

Table A71: Deduplication sensitivity ($16\times$, Normal; five-seed means, %). Dedup+ removes confirmed overlap IDs within each audit scope before training. Δ: Dedup+ minus Full (pp).

Training set	**Avg CEM**	**Avg F1**
Full	41.30	44.50
Dedup+	41.24	44.48
Δ	-0.06	-0.02

A.47 Experimental Configuration and Data Licenses

Table A72: BIRC hyperparameters. Dev: independent 500-query development selection; Eval: benchmark evaluation selection. Sensitivity grids alone do not establish independent selection.

Hyperparameter	Value	Selection basis / grid
Pipeline / interface updates		
τ (CCEF threshold)	0.30	fixed
α (CCEF floor)	0.5	Dev: $\{0.3, 0.5, 0.7\}$
β (ACR sharpness)	10	Dev: $\{5, 10, 20\}$
$\rho_{\min}$	0.25	Dev: $\{0.10, 0.25, 0.40\}$
$\rho_{\max}$	1.0	fixed during selection
w_{cfrs}	0.30	Eval: $0.00, 0.10, 0.20, 0.30, 0.40, 0.50$
γ (IGFR coverage threshold)	0.5	fixed
K_{igfr} (4–64×)	2	Eval; grid: $\{1, 2, 3, 4\}$
K_{igfr} (128×)	1	Eval; grid: $\{1, 2, 3\}$
λ (hidden-state alignment weight)	0.1	measured: $\{0, 0.01, 0.05, 0.1, 0.2, 0.5\}$
ϵ (numerical floor)	10^{-6}	fixed
Ordering temperature (CFRS / CLaRa)	0.1 / 0.02	fixed
Training		
LoRA rank r_{LoRA}	16	matched to CLaRa
LoRA α / dropout / bias	32 / 0.1 / none	fixed
LoRA target modules (BIRC / CLaRa)	`q_proj` / all linear layers	fixed
Learning rate (Mistral/Llama)	2×10^{-4}	fixed
Learning rate (Qwen2.5-14B)	1.6×10^{-4}	0.8× Mistral learning rate
Warmup steps	500	fixed
Optimizer / schedule	AdamW / cosine	fixed
AdamW (β_1, β_2) / ϵ / weight decay	$(0.9, 0.95)$ / 10^{-8} / 0	fixed
Gradient-norm clip	1.0	fixed
Phase-I epochs	3	fixed
Phase-II epochs	5	fixed
Effective batch size (Mistral/Llama)	32	fixed
Effective batch size (Qwen2.5-14B)	16 (gradient accumulation: 2)	fixed for Qwen2.5-14B
Maximum passage-content / query length	768 / 256 tokens	fixed
Phase-I input / target limits	2,048 / 512 tokens	fixed
Phase-II input / target limits	1,024 / 128 tokens	fixed
Answer decoding	greedy (one beam); EOS or 64 new tokens	fixed
Pre-fitting W_{BGE}		
Optimizer / learning rate	AdamW / 5×10^{-4}	fixed
AdamW (β_1, β_2) / ϵ / weight decay	$(0.9, 0.999)$ / 10^{-8} / 0.01	fixed
Pairs / batch / epochs	50,000 / 128 / 2	fixed
Pipeline pool sizes		
AHR pool $\lvert C_1^{\text{AHR}} \rvert$	4,000	fixed
IGFR additions per iteration	at most 200 unseen documents	fixed
MTFRL second-round pool $\lvert D_2 \rvert$	200	fixed
MADS→CCEF	$100 \to 5$	fixed
MTFRL projection-head dimensions		
Mistral-7B/Llama-3-8B	$4096 \to 2048 \to 1024$	backbone hidden size → half-width → BGE output dimension
Qwen2.5-14B	$5120 \to 2560 \to 1024$	backbone hidden size → half-width → BGE output dimension

Table A73: QCA rules: Multi-Hop/Multi-Aspect (26 rules); w_ℓ is rule weight. Entity count uses distinct PERSON/ORG/GPE/LOC/FAC/EVENT/PRODUCT/WORK_OF_ART entities. Contains/starts-with use case-insensitive substring/prefix matching; entity, template, and syntax conditions are separate. Simple rules: Table A74.

Category	Rule ID	w_ℓ	Trigger pattern / condition
Multi-Hop	H01	1.0	entity count ≥ 3
	H02	1.0	contains: "in which country", "in which city", "in which state", "in which year"
	H03	0.9	contains: "born in the same", "headquartered in the", "located in the same"
	H04	0.9	entity count ≥ 2 AND contains: "who invented", "who founded", "who directed", "who wrote", "who discovered"
	H05	0.9	contains: "where was … born", "where did … die", "where is … headquartered"
	H06	0.85	entity count $= 2$ AND contains a bridge connective: "that", "which", "whose", "where"
	H07	0.85	contains: "the {ENTITY} of {ENTITY}" (nested possessive pattern)
	H08	0.8	contains: "same {noun} as", "the same … as"
	H09	0.8	contains: "also known as", "formerly known as", "previously called"
	H10	0.75	subquestion trigger: starts with "what is the … of the {ENTITY} that"
	H11	0.75	contains: "after {ENTITY} {verb}", "before {ENTITY} {verb}"
	H12	0.7	contains: "the {superlative} … in the {ENTITY}" (e.g., "tallest mountain in the country where")
	H13	0.65	entity count $= 2$ AND begins with "what", "who", "when", or "where" AND contains a transitive verb
	H14	0.6	contains: "through which", "via which", "by means of which"
Multi-Aspect	A01	1.0	starts with: "compare", "what are the differences between", "contrast"
	A02	1.0	contains: "advantages and disadvantages", "pros and cons", "benefits and drawbacks"
	A03	0.9	contains: "both … and …", "either … or …" (with two distinct entities)
	A04	0.9	contains: "difference between", "similarities between", "relationship between"
	A05	0.85	contains an enumeration trigger: "list", "enumerate", "name all", "give examples of"
	A06	0.85	contains: "how does … compare to", "how is … different from"
	A07	0.8	contains: "what are the main", "what are the key", "what are the primary" (plural aspect noun)
	A08	0.75	contains: "in terms of", "with respect to", "regarding … and …"
	A09	0.75	contains: "as well as", "in addition to" (with entity count ≥ 2)
	A10	0.7	contains: "multiple {noun}", "various {noun}", "several {noun}" (plural aspect)
	A11	0.65	contains: "what factors", "what aspects", "what characteristics"
	A12	0.6	contains: "overview of", "describe the … of", "explain the {noun} and the {noun}"

Table A74: QCA rules: Simple (12 rules), completing the 38-rule inventory with the notation of Table A73.

Category	Rule ID	w_ℓ	Trigger pattern / condition
Simple	S01	1.0	entity count $= 1$ AND starts with: “what is”, “who is”, “when is”
	S02	1.0	entity count $= 1$ AND starts with: “when was”, “where was”, “how many”
	S03	0.9	entity count $= 1$ AND starts with: “who wrote”, “who invented” (single known entity)
	S04	0.9	starts with: “what year”, “what date”, “in what year”, “how old is”
	S05	0.85	contains: “the capital of”, “the population of”, “the president of” (simple fact)
	S06	0.85	entity count $= 0$ AND contains a copular verb: “is”, “was”, “are”, “were” (definition query)
	S07	0.8	starts with: “what is the name of the”, “who is the current”
	S08	0.8	contains: “the birthplace of”, “the nationality of”, “the height of”
	S09	0.75	starts with: “how tall”, “how long”, “how far”, “how deep”, “how wide”
	S10	0.7	starts with: “what language”, “what currency”, “what color”, “what type of”
	S11	0.65	entity count $= 1$ AND contains neither a bridge connective nor a comparative or superlative
	S12	0.6	fallback (default): no hop or aspect indicators triggered

Note. A query is Multi-Hop when a hop rule fires and the entity count is at least two; it is Multi-Aspect when an aspect rule fires and no hop rule fires; otherwise it is Simple. Hop rules take precedence when both categories match. Rules within each group are ordered by descending w_ℓ. Brace-delimited placeholders and ellipses denote declarative pattern families. The w_ℓ values were held fixed throughout evaluation.

Statistical protocol. The five seeds are 42, 123, 456, 789, and 2024. Statistics are computed before display rounding. Native-pipeline seed tables report sample standard deviations and two-sided 95% Student-t intervals with four degrees of freedom. Paired seed-level comparisons use the differences between checkpoints with matching seeds. For each benchmark, paired tests average the five seed-aligned differences per evaluation item and apply a 10,000-resample paired query bootstrap (Berg-Kirkpatrick et al., 2012). Avg contrasts resample within each benchmark before taking the unweighted four-benchmark mean. These query-level intervals and p-values are conditional on the evaluation panels and five aligned checkpoint pairs; seed-level uncertainty is summarized separately by the reported Student-t intervals. Holm adjustment and its comparison family are stated where applied. Paired query-level bootstrap p-values are otherwise unadjusted.

The datasets and retrieval corpora used in Phase II and evaluation have the following licenses or access terms: NQ (CC BY-SA 3.0), HotpotQA (CC BY-SA 4.0), MuSiQue (CC BY 4.0), 2WikiMultiHopQA (Apache 2.0), and KILT’s 2019-08-01 Wikipedia text under the applicable Wikipedia reuse terms (principally CC BY-SA 3.0).

Table A72 consolidates BIRC’s configuration; Appendix A.29 provides the corresponding CLaRa configuration. The statistical protocol above expands the summary in §4; responsible-research details appear in the Ethics Statement.

A.48 Extended Limitations and Scope

Fixed feedback depth. The primary interface uses one memory-conditioned feedback update during both training and evaluation, assigning the same feedback budget to every query. Passage selection remains discrete. The two- and three-round configurations in Appendix A.25 apply the operator repeatedly at test time while keeping its parameters fixed.

Soft memory weighting. ACR scales memory activations while retaining all base positions. Physical tensor and KV-cache lengths therefore depend on the base memory counts. First-pass prefix counts determine MTFRL pooling; final-set prefix counts summarize the gates applied to reader input. The primary configuration uses one global floor across compression ratios, so the same score-to-scale rule operates at different memory budgets. The physical-allocation and adaptive-floor controls evaluate separate changes to these design choices.

Fidelity signal and gradient surrogates. CFRS measures source predictability conditioned on latent memory and preceding source tokens under a frozen decoder. This reconstruction criterion supplies an indirect signal for evidence ordering; retrieval relevance supplies the query-dependent signal. Reconstruction fidelity does not directly score whether a passage supports the answer. During training, straight-through surrogates carry QA gradients through selected-document scores and fidelity, while forward selection and ordering remain discrete.

Retrieval dependence and evaluation scope. The supporting retriever uses a frozen BGE embedding space and corpus index, MADS–CCEF scoring of the combined candidates, and English surface rules for routing. These components determine which evidence reaches the compressor. The evaluations cover four English Wikipedia-based QA benchmarks, 7–14B trainable reader backbones, and a frozen 4-bit-quantized Llama-3-70B-Instruct reader with a 1,024-token input cap. Native-pipeline comparisons evaluate complete systems. Shared-candidate, selector-input, factorial, text-feedback, and physical-allocation controls evaluate the specified interventions within this implementation. In particular, shared candidate retraining is available for CLaRa, but not for SARA, SeleCom, or OSCAR; native-pipeline advantages over the latter do not isolate the compressor. There is no matched accuracy–latency evaluation against xRAG or MemoRAG.

Lexical detail and retrieval robustness. Mean QA and reconstruction scores do not establish exact retention of rare entity names, numbers, dates, or negation. Popularity-stratified evaluation on PopQA or EntityQuestions (Mallen et al., 2023) would separate head-entity performance from long-tail performance. No such stratification or controlled counterfactual-document evaluation is reported here. Reconstruction can favor predictable but unhelpful text, and uniform outer pooling can propagate an irrelevant first-pass passage into the feedback direction.

Selection bias and data provenance. The IGFR caps and CFRS weight were selected on the benchmark evaluation outcomes. Repeated seeds and paired bootstrap tests quantify uncertainty conditional on this choice; they do not remove the resulting optimism. Selection on development data followed by evaluation on untouched questions is required for confirmatory claims. The original 50,000-pair projection-fit manifest is needed to resolve its dataset and split composition. For Phase I, exact URL exclusion is exhaustive, while semantic overlap review covers a category-balanced 200,000-row sample. Neither that sample nor the associated Dedup+ control establishes full-corpus absence of answer or reasoning overlap.

A.49 Mechanistic Interpretation and Failure Modes

This section analyzes the evaluated equations and implementation. The alternatives discussed below are possible controls, not additional measured variants.

A.49.1 Input gates in a pre-normalized reader

Let $v(\boldsymbol{m}) = d_h^{-1}\|\boldsymbol{m}\|_2^2$ and let $\delta > 0$ be the reader's RMSNorm numerical constant, distinct from the score-normalization ϵ. For a positive scalar gate g and learned elementwise gain $\boldsymbol{a}$, the RMSNorm

definition (Zhang & Sennrich, 2019) gives

$$\mathrm{RMSNorm}(g\boldsymbol{m}) = \frac{g\boldsymbol{a} \odot \boldsymbol{m}}{\sqrt{g^2 v(\boldsymbol{m}) + \delta}} = \frac{\boldsymbol{a} \odot \boldsymbol{m}}{\sqrt{v(\boldsymbol{m}) + \delta/g^2}}. \tag{25}$$

When $g^2 v(\boldsymbol{m}) \gg \delta$, this is approximately independent of g. With $\delta = 0$ it is exactly scale invariant for nonzero $\boldsymbol{m}$ and $g > 0$. Thus an ordinary positive gate need not attenuate the first attention layer's queries, keys, or values in proportion to its magnitude.

Three routes still expose the gate. First, a pre-normalized residual update at memory position t has the form $\boldsymbol{x}_t^+ = g_t \boldsymbol{m}_t + \mathcal{A}_t(\mathrm{RMSNorm}(\boldsymbol{X}))$, where $\boldsymbol{X}$ contains the full gated input sequence and $\mathcal{A}$ is the attention block. Even if its normalized inputs are unchanged, the relative weight of the residual and attention output changes. For non-collinear inputs and updates, normalization before the MLP and later layers can then change direction. This does not guarantee reduced attention to a low-gate position. Second, when $g^2 v(\boldsymbol{m})$ approaches δ, the numerical constant matters; for fixed $\delta > 0$, the normalized state tends to zero as $g \to 0$. Third, MTFRL pools the gated states before reader normalization, with relevance-dependent prefix membership. These mechanisms show why gating can affect the full system, while leaving its layerwise contribution an empirical question. All positions remain in the attention mask and KV cache.

Boundary and floor conventions. The cutoff is a sigmoid midpoint: $g_i^{(t)} = 1/2$ when $t = \rho_i K_i^{(0)}$. In particular, a hypothetical exact $\rho_i = 1$ halves the final state; with Eq. 3's positive ϵ, the largest realized scale is slightly below one for a finite score span. An ungated state and a maximally scaled ACR state are therefore distinct. The evaluated implementation uses this convention, without a half-position shift. Equation 5 uses floor, not nearest-integer rounding, and enforces at least one pooled position even when the soft cutoff is below one.

Local score normalization. For score span $b = \max_i s_i - \min_i s_i$, the spread of ACR scales is $(\rho_{\max} - \rho_{\min})b/(b + \epsilon)$. Thus a narrow span can still produce nearly the full gate range when $b \gg 10^{-6}$; only spans comparable to the numerical floor are substantially attenuated. This rule encodes relative rank-set contrast, not absolute relevance. Equal scores place all passages at $\rho_{\min}$. CFRS has the same relative-scale sensitivity, and its detached normalization yields the factor $1/(b_f+\epsilon)$ in the error gradient, where b_f is the fidelity span. Bounded forward scores do not bound this factor away from large values. The configured gradient-norm clip of 1.0 limits the total optimizer gradient, but does not calibrate either score. Absolute-score calibration or shrinkage toward a uniform gate would change the method and requires separate validation.

A.49.2 Passage pooling and retrieval-space alignment

Writing $\boldsymbol{v}_i = T_i^{-1} \sum_{t \le T_i} g_i^{(t)} \boldsymbol{m}_i^{(t)}$ makes Eq. 6's outer mean explicit. Dividing by T_i prevents a passage from receiving more weight solely because more of its positions are pooled. However, every passage has coefficient $1/5$ regardless of retrieval score. The vectors can differ in norm and direction, so equal coefficients do not imply exactly equal influence on the normalized query. Nor do the inner gates guarantee that an irrelevant passage is suppressed: most positions within its short prefix can still have gates close to one.

A score-weighted alternative is $\mathrm{Normalize}(\sum_i a_i \boldsymbol{v}_i)$, with $a_i = \exp(s_i/\tau_p)/\sum_j \exp(s_j/\tau_p)$ and $\tau_p > 0$. It could reduce weak-candidate influence, but could also reinforce an erroneous first-pass ranking; the reported results do not compare these choices. Noise and counterfactual controls should replace known first-pass passages while holding the question and checkpoint fixed, and measure both final supporting-page recall and answer quality, alongside feedback-query displacement.

The pre-fitted W_{BGE} aligns base-model question states to indexed passage embeddings, whereas P_{fb} consumes pooled compressor states. Its exact W_{BGE} initialization is therefore a coordinate initialization, not a proof that these two input distributions are aligned. Phase II supplies selected-document feedback gradients; hard search and original-query union rescoring still restrict which retrieved evidence can train this projection.

A.49.3 FIDELITY, SORTING GRADIENTS, AND RECONSTRUCTION COST

Fidelity measures how well the frozen decoder predicts a source conditioned on its gated memory and the preceding gold source tokens. It can reflect retained detail, but also the decoder's prior and the source's predictability. It does not isolate information contributed by memory, and it is not an answer-support or exact-token-preservation metric. Combining it with retrieval relevance is therefore a heuristic whose support comes from the stated controls.

Gradient through the ordering surrogate. Let $a_r = \mathrm{sg}(u_{\pi_r})$ and $\ell_{ri} = -(u_i - a_r)^2/\tau_{\mathrm{ord}}$. Treating anchors as constants,

$$\frac{\partial P^{\mathrm{soft}}_{ri}}{\partial u_k} = P^{\mathrm{soft}}_{ri}\big(\mathbf{1}[i=k] - P^{\mathrm{soft}}_{rk}\big)\frac{-2(u_k - a_r)}{\tau_{\mathrm{ord}}}. \tag{26}$$

For $k = \pi_r$ this derivative is zero in row r; that score can still receive gradients through other rows whose anchors differ. Exact score ties can eliminate those paths too. The row-wise assignment is not constrained to be doubly stochastic, and its derivative is a biased surrogate, not a derivative of the hard sort. The forward permutation remains exact.

With $\Delta_f = \mathrm{sg}(\max_j(-e_j) - \min_j(-e_j) + \epsilon)$, the CFRS contribution to the QA derivative is

$$\left.\frac{\partial \mathcal{L}_{\mathrm{QA}}}{\partial e_i}\right|_{\mathrm{CFRS}} = -\frac{w_{\mathrm{cfrs}}}{\Delta_f}\frac{\partial \mathcal{L}_{\mathrm{QA}}}{\partial u_i}. \tag{27}$$

If the local QA gradient favors lowering u_i, its contribution can increase e_i. There is no demotion-specific stop-gradient and no separate Phase-II reconstruction-preservation term. Although ACR conditions the memory on query scores, the underlying compressor is query-agnostic, so updates induced by one query can conflict with faithful encoding for another. Phase-I pretraining, frozen decoder weights, and gradient clipping do not rule out that conflict. The CFRS-free control removes the ordering operation as well as its gradient; it does not isolate this training effect. A forward-identical control with detached fidelity and a reconstruction audit would be needed for that purpose.

Teacher-forced computation. The five selected passages require five reconstruction-only passes, at most 768 source target tokens each, with logit processing in chunks of 128. Each pass receives the complete shifted source, allowing target positions to be scored in one teacher-forced forward invocation rather than generating them autoregressively. This does not remove the decoder's source-token computation or its temporary activation cost. Reconstruction temporaries can be released before answer generation at inference; the answer reader then uses the latent prompt. The reported 2.68-second compression/CFRS category at $16\times$ includes this work but does not separately time CFRS. Its cost should not be inferred from the memory-token count alone.

A.49.4 SYMMETRIC HIDDEN-STATE ALIGNMENT

For pooled states $\boldsymbol{m}$ and $\boldsymbol{h}$, the HSA gradients are $2(\boldsymbol{m} - \boldsymbol{h})$ and $2(\boldsymbol{h} - \boldsymbol{m})$, respectively. Both move toward each other; the query–answer branch is not a context-rich fixed target. HSA alone admits identical-state solutions, while the QA loss requires useful output predictions. Neither the frozen base nor the combined objective proves that all pooled representations remain informative. Table A11 directly evaluates detaching the query–answer mean: its $+0.05$-point Avg F1 difference is descriptive and does not establish a reliable improvement or certify the absence of collapse. Monitoring representation variance and effective rank, jointly with answer quality, would address a different question from these endpoint controls.