# CLOSING THE GAP: A TWO-STAGE TRAINING RECIPE FOR SOFT CONTEXT COMPRESSION IN RAG

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Retrieval-Augmented Generation (RAG) improves knowledge-intensive generation by conditioning language models on retrieved documents, but processing these documents becomes increasingly expensive as retrieval depth grows. Soft context compression reduces this cost by encoding documents into compact continuous representations that can be precomputed and reused across queries. However, many existing methods train compressed models by distilling from a full-context teacher. When the teacher is wrong, such distillation can reinforce its errors, while teacher imitation provides no direct signal for improving beyond the teacher. We propose DEX-Comp, a two-stage training recipe that separates reliable imitation from targeted exploration. Pure Distillation learns only from teacher-correct questions to mitigate error propagation, while Hard Exploration applies outcome-based reinforcement learning to teacher-failed questions to directly optimize answer correctness. Across five open-domain QA benchmarks and retrieval depths from top-5 to top-30, DEX-Comp at $16\times$ compression outperforms all evaluated compression baselines and surpasses the untuned full-context RAG model in average accuracy, while reducing time-to-first-token by $4.4\times$–$23.7\times$. Evaluations across additional datasets and backbones further demonstrate its generalization.

## 1 INTRODUCTION

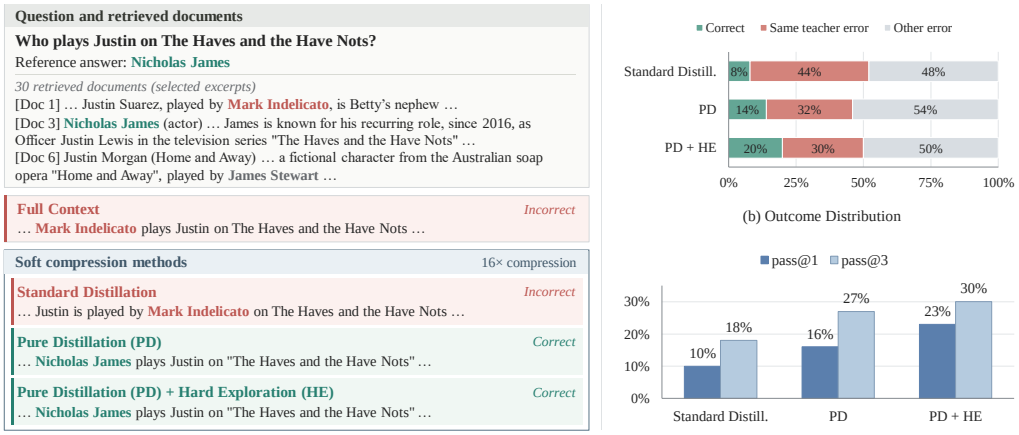
Retrieval-Augmented Generation (RAG) improves knowledge-intensive generation by conditioning large language models (LLMs) on retrieved documents (Gao et al., 2023). However, expanding the retrieved context introduces two challenges: (1) Longer inputs increase computation and memory requirements, limiting the efficiency of generation (Chevalier et al., 2023); and (2) retrieval may introduce redundant or irrelevant information that distracts the model and degrades generation quality (Jiang et al., 2024). These challenges motivate context compression: reducing the retrieved context before it is consumed by the language model.

Existing approaches broadly fall into hard context compression (Yun & Kim, 2026; Zhao et al., 2026b; Shi et al., 2026; Mohamed et al., 2026; Tang et al., 2025; Zhao et al., 2025b; Fei et al., 2025; Hu et al., 2025) and soft context compression (Tang et al., 2026a;b;c;d; Zhao et al., 2026a; Liu et al., 2026a; Louis et al., 2025b; Zhao et al., 2025a; Li et al., 2025). Hard compression reduces context by extracting salient spans (Jiang et al., 2023b) or summarizing documents into shorter text (Jin et al., 2025a), but aggressive compression can discard information useful for downstream generation (Lajewska et al., 2025). Soft compression instead encodes each document into a much shorter sequence of continuous embeddings (Zhang et al., 2025). Because these embeddings need not correspond to readable tokens, they provide a flexible learned representation of the original document (Kuratov et al., 2025). In this work, we focus on *query-independent soft compression*, where documents are compressed without access to the query (Liu et al., 2025b; Dai et al., 2025). Their compressed representations can therefore be precomputed offline and reused across queries, substantially reducing the context processed by the decoder at inference time.

A central challenge in soft compression is how to train these compressed representations. A common pipeline first pretrains the compressor on unlabeled text through reconstruction or next-token prediction and then fine-tunes it for downstream RAG question answering (QA) (Ge et al., 2024; Rau et al., 2024b). For QA fine-tuning, distillation from an uncompressed full-context model has been shown to outperform direct supervision with reference answers, possibly because short reference answers



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.




(a) Response Comparison

(b) Outcome Distribution

(c) Pass@1 and Pass@3

Figure 1: **(a) Representative case.** Standard distillation reproduces an incorrect answer from the full-context teacher, whereas Pure Distillation (PD) and PD with Hard Exploration (PD+HE) answer correctly. **(b) Outcome distribution.** Under greedy decoding on a fixed subset of teacher-failed questions, PD reduces the frequency of reproducing teacher errors, while HE further increases correct answers. **(c) Sampling performance.** On the same questions, PD improves pass@1 and pass@3 over standard distillation, with further gains from HE.

are poorly aligned with the model's response distribution (Cheng et al., 2024; Guo et al., 2025). More recently, Louis et al. (2025a) showed that distillation alone can achieve strong performance without separate compression pretraining. Together, these results establish the full-context model as an effective source of supervision for learning soft compressed representations.

However, an effective teacher is not necessarily a reliable oracle. This creates two limitations for distillation-based compression training. (1) *Imitation of teacher errors.* Standard distillation approaches imitate the full-context teacher regardless of whether its response is correct. When the teacher fails, this supervision can propagate the same error to the compressed model. Figure 1(a) shows a representative example: despite the correct evidence appearing in the retrieved documents, the full-context teacher produces an incorrect answer, which is subsequently reproduced by standard distillation. More broadly, Figure 1(b) shows that a substantial fraction of errors made by the distilled model match those of the teacher. (2) *Limited improvement beyond imitation.* Distillation optimizes agreement with the teacher rather than answer correctness itself. It therefore provides no direct learning signal for correcting cases in which the teacher fails, placing a natural ceiling on what can be achieved through imitation alone.

This observation suggests that teacher-correct and teacher-failed examples should play different roles in compression training. When the full-context teacher answers correctly, its behavior provides useful supervision for learning how to answer from compressed representations. When it answers incorrectly, however, imitating that behavior is undesirable; these examples instead provide an opportunity to optimize the compressed model directly against task outcomes. This naturally raises two research questions: *(1) How can we use a full-context teacher to reliably initialize a compressed model without propagating its observed errors? (2) How can we directly optimize the compressed model for answer correctness beyond imitation?*

To address these questions, we propose **DEX-Comp** (**D**istillation and **Ex**ploration for soft **Comp**ression), a two-stage training recipe that assigns different learning signals according to teacher correctness. In the first stage, *Pure Distillation (PD)*, we distill only on teacher-correct questions, providing a reliable initialization while avoiding direct imitation of the teacher's observed errors (see Figure 1(b,c)). In the second stage, *Hard Exploration (HE)*, we apply outcome-based reinforcement learning to teacher-failed questions, rewarding answer correctness rather than agreement with the teacher. HE therefore turns examples with unreliable imitation targets into opportunities for direct task optimization under compressed inputs, enabling further improvement beyond teacher imitation (see Figure 1(b,c) and Section 6.1).

This separation is particularly attractive for soft compression. Rather than performing reinforcement learning over the full retrieved context, DEX-Comp performs reinforcement learning over substantially shorter compressed representations. As a result, it improves over an untuned full-context RAG model at substantially lower training cost than full-context RL and remains feasible at retrieval depths where full-context RL exceeds the available memory budget (Section 6.2). Thus, compression serves not only as an inference-time efficiency mechanism, but also as an efficient substrate for post-training.

Our contributions are threefold: (1) We identify two limitations of indiscriminate teacher imitation for soft-compression training: distillation can propagate teacher errors, while imitation alone provides no direct mechanism for correcting them. (2) We introduce DEX-Comp, which assigns different learning signals according to teacher correctness, using reliable distillation on teacher-correct examples and outcome-based reinforcement learning on teacher-failed examples. (3) Across five open-domain QA benchmarks and retrieval depths from top-5 to top-30, DEX-Comp at $16\times$ compression outperforms all evaluated compression baselines and surpasses the untuned full-context RAG model in average accuracy, while reducing time-to-first-token by $4.4\times$–$23.7\times$. Additional evaluations across datasets and model backbones further demonstrate its generalization.

## 2 Related Work

**Retrieval-Augmented Generation** RAG augments language models with external knowledge and has been widely studied for knowledge-intensive tasks (Guan et al., 2026; Wei et al., 2025; Gutiérrez et al., 2025; Shi et al., 2024; Min et al., 2023). Prior work improves RAG through joint retriever–generator training, retriever alignment, and adaptive retrieval and generation (Guu et al., 2020; Lin et al., 2024; Asai et al., 2024), while a complementary line of work studies the efficiency of processing increasingly large retrieved contexts.

**Context Compression** RAG efficiency can be improved through inference-time KV compression (Hao et al., 2026; 2025; Li et al., 2024a; Chang et al., 2025) or pre-inference context compression (Li et al., 2024b). The latter includes hard compression, which extracts or summarizes textual content (Zhao et al., 2025c; Chung et al., 2024; Hwang et al., 2025; Chirkova et al., 2025; Li et al., 2023; Li, 2023; Xu et al., 2024; Liu et al., 2025a; Fang et al., 2025; Jin et al., 2025a; Chuang et al., 2024), and soft compression, which encodes context into continuous embeddings (Feng et al., 2026). We focus on query-independent soft compression, where documents are encoded offline without access to the query, unlike query-dependent methods that condition compression on the query (Liu et al., 2026b; Liskavets et al., 2025; Chen et al., 2025; Cui et al., 2025; Wang et al., 2023; Yoon et al., 2024).

A common soft-compression pipeline first pretrains a compressor on unlabeled text and then fine-tunes it on downstream QA with reference-answer supervision (Jin et al., 2025b; Bulatov et al., 2022; Wingate et al., 2022; Qin et al., 2024; Mu et al., 2023; Tan et al., 2024). Other methods use distillation from the uncompressed model (Guo et al., 2025; Feldman & Artzi, 2025; Cheng et al., 2024; Yen et al., 2024), with Louis et al. (2025a) showing that distillation alone can remove separate pretraining. Ting et al. (2026) instead combine pretraining and reference-answer fine-tuning with reinforcement learning. *Existing methods therefore combine pretraining, teacher imitation, and outcome optimization in different ways, but do not explicitly distinguish when teacher imitation is reliable and when direct outcome optimization is preferable. DEX-Comp makes this distinction explicit by using teacher-correct examples for distillation and teacher-failed examples for exploration.*

## 3 Method

**Problem Formulation.** A standard RAG system consists of a retriever $\mathcal{R}$ and a decoder $\mathcal{T}$ (Karpukhin et al., 2020). Given a query $q$, the retriever returns the top-$k$ documents $\mathbf{D}_q = \{d_1, \ldots, d_k\}$ from a corpus $\mathbb{C}$, and the decoder generates a response conditioned on the retrieved documents:

$$r \sim \mathcal{T}(\cdot \mid \mathbf{D}_q, q).$$

Processing the retrieved documents in full becomes increasingly expensive as retrieval depth grows. Soft context compression reduces this cost by using a compressor $\mathcal{C}_\phi$ to replace each document $d_i$ with

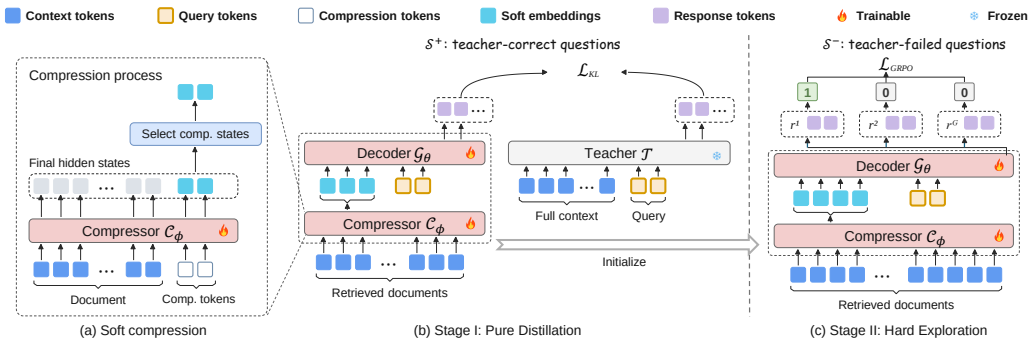



Figure 2: Overview of DEX-Comp. **(a) Soft compression.** A compressor turns each document into a shorter sequence of soft embeddings by selecting the final-layer hidden states at compression-token positions. The embeddings can be precomputed offline and consumed by the decoder at inference time. **(b) Pure Distillation.** On teacher-correct queries ($\mathcal{S}^+$), KL distillation aligns the compressed model's response distribution with that of a frozen RAG teacher reading the full retrieved documents. **(c) Hard Exploration.** Initialized from Stage I, the compressed model trains on teacher-failed queries ($\mathcal{S}^-$) using outcome-based GRPO, directly rewarding answer correctness rather than teacher agreement.

a shorter sequence of continuous embeddings $\mathbf{e}_i = \mathcal{C}_\phi(d_i)$, which are consumed by a compression-aware decoder $\mathcal{G}_\theta$. These continuous representations have no explicit supervision target, making it non-trivial to jointly train the compressor and decoder. Prior work therefore commonly uses an uncompressed full-context model as a teacher (Cheng et al., 2024; Louis et al., 2025a). As discussed in Section 1, such a teacher provides useful supervision when it answers correctly, but becomes an unreliable imitation target when it fails.

Motivated by this distinction, we propose **DEX-Comp** (**D**istillation and **Ex**ploration for soft **Comp**ression), a two-stage training recipe that assigns different learning signals according to teacher correctness. *Pure Distillation (PD)* uses teacher-correct examples to provide a reliable initialization, while *Hard Exploration (HE)* applies outcome-based reinforcement learning to teacher-failed examples to directly optimize answer correctness. We first describe the compression architecture (§3.1) and then introduce the two training stages (§3.2–§3.3).

### 3.1 DEX-Comp Architecture

DEX-Comp consists of a compressor $\mathcal{C}_\phi$ and a decoder $\mathcal{G}_\theta$. For each document $d_i$, the input text is concatenated with $m_i = \lfloor L_i/\tau \rfloor$ special compression tokens, where $L_i$ is the document length and $\tau$ is a fixed compression rate. The full sequence is encoded by $\mathcal{C}_\phi$, and the final-layer hidden states at the compression-token positions form the compressed embedding sequence

$$\mathbf{e}_i = \mathcal{C}_\phi(d_i) = \left(e_i^1, \ldots, e_i^{m_i}\right).$$

The entire compression process is run offline once per document in the corpus $\mathbb{C}$, with all $\mathbf{e}_i$ precomputed and cached for instant access at inference time. The decoder is an autoregressive language model. Given the cached embeddings $\{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$ of the top-$k$ retrieved documents and a query $q$, $\mathcal{G}_\theta$ generates the response token by token from

$$P_\theta(r_t \mid r_{<t}, q, \mathbf{e}_1, \ldots, \mathbf{e}_k).$$

### 3.2 Training Stage I: Pure Distillation

Distillation provides an effective way to initialize the jointly trained compressor and decoder, but the full-context teacher is not correct on every training example. We therefore restrict imitation to examples for which the teacher provides a reliable demonstration. Specifically, we partition the training set according to whether the teacher's response matches the gold answer:

$$\mathcal{S}^+ = \left\{(q, \mathbf{D}_q) : \mathrm{eval}(\hat{r}_\mathcal{T}, r^\star) = 1\right\}, \qquad \mathcal{S}^- = \left\{(q, \mathbf{D}_q) : \mathrm{eval}(\hat{r}_\mathcal{T}, r^\star) = 0\right\}, \tag{1}$$

where $\hat{r}_\mathcal{T} \sim \mathcal{T}(\cdot \mid \mathbf{D}_q, q)$ is the teacher response and $\mathrm{eval}(\cdot,\cdot) \in \{0, 1\}$ checks whether the prediction matches the gold answer $r^\star$. Specifically, we use Gemini 3 Flash to perform this check.

PD operates only on the teacher-correct subset $\mathcal{S}^+$, while $\mathcal{S}^-$ is reserved for HE. For each $(q, \mathbf{D}_q) \in \mathcal{S}^+$, the teacher reads the original retrieved tokens while the student reads their compressed counterparts $\mathbf{e}_i = \mathcal{C}_\phi(d_i)$, and we minimize the Kullback–Leibler divergence between the two resulting response distributions:

$$\mathcal{L}_{\text{PD}}(\phi, \theta) = \mathbb{E}_{(q,\mathbf{D}_q)\sim\mathcal{S}^+}\Big[D_{\text{KL}}\big(P_{\mathcal{T}}(r \mid \mathbf{D}_q, q) \,\big\|\, P_\theta(r \mid \mathbf{e}_1, \ldots, \mathbf{e}_k, q)\big)\Big]. \tag{2}$$

Notably, the gradient flows through both $\theta$ and, via $\mathbf{e}_i$, through $\phi$.

**Why filter teacher-failed examples?** Teacher-correct responses provide reliable demonstrations for learning through compressed representations, whereas imitating teacher-failed responses directly reinforces undesirable behavior. We therefore reserve the latter for outcome optimization in Stage II.

### 3.3 Training Stage II: Hard Exploration

Pure Distillation provides a reliable initialization but learns only from teacher-correct behavior. For $\mathcal{S}^-$, neither the incorrect teacher responses nor direct fine-tuning on short reference answers provides a desirable imitation target, as the latter can distort the model's response distribution (Guo et al., 2025). We therefore use reference answers as outcome signals rather than generation targets. Starting from the PD checkpoint, Hard Exploration (HE) trains on $\mathcal{S}^-$ to directly optimize answer correctness.

Specifically, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each query $q \in \mathcal{S}^-$, we draw a group of $G$ candidate responses from the current student policy $\pi_{\phi,\theta}$:

$$r^{(g)} \sim \pi_{\phi,\theta}\big(\cdot \mid \mathbf{e}_1, \ldots, \mathbf{e}_k, q\big), \qquad \mathbf{e}_i = \mathcal{C}_\phi(d_i), \qquad g = 1, \ldots, G, \tag{3}$$

and assign each rollout a binary outcome reward

$$R^{(g)} = \text{eval}(r^{(g)}, r^\star) \in \{0, 1\}, \tag{4}$$

Specifically, we adopt containment exact match (CEM) as the reward: a rollout receives 1 if it contains the gold answer and 0 otherwise. We cap each rollout at 32 generated tokens to constrain response length. Group-relative advantages are then formed by standardizing rewards within each group:

$$A^{(g)} = \frac{R^{(g)} - \mu_R}{\sigma_R + \epsilon}, \qquad \mu_R = \frac{1}{G}\sum_{g=1}^{G} R^{(g)}, \qquad \sigma_R^2 = \frac{1}{G}\sum_{g=1}^{G}\big(R^{(g)} - \mu_R\big)^2. \tag{5}$$

The student is then updated with the clipped policy-gradient surrogate, regularized toward the PD checkpoint $\pi_{\text{ref}}$:

$$\mathcal{L}_{\text{HE}}(\phi, \theta) \;=\; -\mathbb{E}_{q\sim\mathcal{S}^-}\left[\frac{1}{G}\sum_{g=1}^{G}\frac{1}{|r^{(g)}|}\sum_{t=1}^{|r^{(g)}|}\min\Big(\rho_t^{(g)}A^{(g)}, \text{clip}\big(\rho_t^{(g)}, 1-\varepsilon, 1+\varepsilon\big)A^{(g)}\Big)\right] \tag{6}$$
$$+\beta\, D_{\text{KL}}[\pi_{\phi,\theta} \,\|\, \pi_{\text{ref}}],$$

where

$$\rho_t^{(g)} = \frac{\pi_{\phi,\theta}(r_t^{(g)} \mid r_{<t}^{(g)}, q, \mathbf{e}_{1:k})}{\pi_{\text{old}}(r_t^{(g)} \mid r_{<t}^{(g)}, q, \mathbf{e}_{1:k})}, \tag{7}$$

$\varepsilon$ is the clipping range, and $\beta$ controls the KL regularization.

Restricting HE to $\mathcal{S}^-$ is deliberate: these are precisely the examples for which teacher imitation provides an unreliable learning signal. Rather than reproducing the teacher's failed behavior, HE directly rewards successful responses generated from compressed inputs. Together, PD and HE assign imitation and outcome optimization to the subsets where each learning signal is most appropriate.

## 4 Experimental Setup

### 4.1 Datasets and Evaluation

We train and evaluate on five open-domain QA datasets: Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) for factoid QA, HotpotQA (Yang et al., 2018) for multi-hop

QA, ASQA (Stelmakh et al., 2022) for ambiguous long-form QA, and PopQA (Mallen et al., 2023) for long-tail QA. We use the Wikipedia-KILT corpus preprocessed by Rau et al. (2024b) for retrieval.

We evaluate retrieval depths from top-5 to top-30 using two metrics: (i) **Containment Exact Match (CEM)**, which checks whether a ground-truth answer appears in the prediction (Cheng et al., 2024; Louis et al., 2025a); and (ii) **LLM-as-judge**, where Gemini 3 Flash[1] judges prediction correctness given the ground-truth answers. The judging prompt is provided in Appendix A.

### 4.2 Baseline Methods

We compare DEX-Comp against two groups of baselines. (i) **Uncompressed RAG.** This baseline feeds the full retrieved context into the language model without compression, serving as the reference for measuring the compression gap. (ii) **Query-independent compression methods.** Like DEX-Comp, these methods compress documents without access to the query, allowing document representations to be computed independently of individual queries. We include LLMLingua-2 (Pan et al., 2024) as a hard-compression baseline, alongside the soft-compression methods xRAG (Cheng et al., 2024), ICAE (Ge et al., 2024), COCOM (Rau et al., 2024b), SAC (Liu et al., 2025b), RL-Comp (Ting et al., 2026), and PISCO (Louis et al., 2025a). To improve comparability, all methods use Mistral-7B (Jiang et al., 2023a) as the decoder backbone.

### 4.3 Implementation Details

We use BERGEN (Rau et al., 2024a) for preprocessing and retrieval, with Splade-v3 (Lassance et al., 2024) as the retriever and DeBERTa-v3 (He et al., 2023) as the reranker. All methods use Mistral-7B-Instruct (Jiang et al., 2023a) as the backbone LLM with temperature 0. For DEX-Comp, the compressor and decoder share the backbone and use separate LoRA (Hu et al., 2022) adapters.

We combine the training splits of all five datasets into a unified training pool and generate a full-context teacher response for each question. Using Gemini 3 Flash, we partition the pool into teacher-correct $\mathcal{S}^+$ and teacher-failed $\mathcal{S}^-$ examples as defined in Eq. 1. PD trains on $\mathcal{S}^+$, and HE initializes from PD and trains on $\mathcal{S}^-$. We train separate variants across retrieval depths and compression rates; unless otherwise specified, experiments use $16\times$ compression with matched training and inference depths. All experiments run on 8 RTX PRO 6000 Blackwell GPUs using PyTorch and Transformers.

## 5 Experimental Results

### 5.1 Main Results

With matched training and inference depths (Table 1), DEX-Comp at $16\times$ compression surpasses uncompressed RAG on every dataset and depth, with average CEM / LLM gains above two points ($p < 0.05$, paired permutation test). To our knowledge, this is the first query-independent soft compression method shown to surpass the uncompressed RAG model across multiple realistic retrieval depths, establishing a new state of the art among comparable compression methods.

### 5.2 Inference Efficiency

On a single RTX PRO 6000 Blackwell GPU with batch size 8, DEX-Comp improves Time-To-First-Token (TTFT), GFLOPs, and peak memory over full-context RAG, with larger gains at greater retrieval depths (Table 2). Offline compression and embedding storage incur a one-time corpus-level cost amortized across queries; only updated documents require recompression.

## 6 Analysis

### 6.1 Ablation Study

Table 3 validates both stages of DEX-Comp and the use of teacher correctness to allocate different learning signals. PD improves average LLM-judged accuracy from 64.79 to 67.75 before RL,

[1] https://ai.google.dev/gemini-api/docs/models/gemini-3-flash-preview

Table 1: Results with Mistral-7B at top-$5/15/30$. Best and second-best scores at each depth are bold and underlined. †: retrained at the corresponding top-$k$ because matching checkpoints are unavailable.

| Method | Comp. Rate | NQ | | TriviaQA | | HotpotQA | | ASQA | | PopQA | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CEM | LLM | CEM | LLM | CEM | LLM | CEM | LLM | CEM | LLM | CEM | LLM |
| **Top-5** | | | | | | | | | | | | | |
| Full context | - | 58.65 | 73.67 | 90.18 | 89.04 | 46.45 | 55.57 | 68.99 | 77.00 | 58.53 | 57.36 | 64.56 | 70.53 |
| LLMLingua-2 | 4× | 49.74 | 63.59 | 84.41 | 83.19 | 36.43 | 45.50 | 59.49 | 65.30 | 39.07 | 39.70 | 53.83 | 59.46 |
| xRAG | 128× | 37.36 | 51.46 | 78.06 | 78.54 | 27.66 | 35.27 | 43.25 | 50.63 | 28.15 | 29.12 | 42.90 | 49.00 |
| ICAE | 4× | 41.59 | 56.33 | 77.78 | 79.27 | 28.75 | 40.48 | 47.68 | 60.65 | 38.28 | 40.24 | 46.82 | 55.39 |
| COCOM | 16× | 30.35 | 45.05 | 67.99 | 71.22 | 23.25 | 34.71 | 37.87 | 48.31 | 19.74 | 22.37 | 35.84 | 44.33 |
| SAC† | 16× | 51.02 | 65.85 | 86.43 | 85.67 | 41.34 | 50.38 | 59.52 | 66.17 | 41.49 | 42.05 | 55.96 | 62.02 |
| RLComp† | 16× | 56.92 | 71.63 | 88.61 | 87.15 | 44.33 | 53.04 | 57.07 | 64.29 | 38.20 | 38.56 | 57.03 | 62.93 |
| PISCO | 16× | 56.61 | 70.81 | 88.01 | 87.06 | 43.88 | 51.80 | 66.24 | 74.26 | 55.95 | 54.78 | 62.14 | 67.74 |
| DEX-Comp | 128× | 59.29 | 73.46 | 91.32 | 87.92 | 44.52 | 53.82 | 69.51 | 77.64 | 50.80 | 49.63 | 63.09 | 68.49 |
| DEX-Comp | 16× | **62.71** | **76.52** | **92.72** | **89.97** | **51.04** | **57.55** | **75.21** | **79.96** | **60.22** | **58.70** | **68.38** | **72.54** |
| **Top-15** | | | | | | | | | | | | | |
| Full context | - | 58.34 | 73.49 | 89.65 | 89.26 | 45.79 | 54.91 | 68.67 | 78.16 | 57.24 | 56.88 | 63.94 | 70.54 |
| LLMLingua-2 | 4× | 50.30 | 65.00 | 84.71 | 82.70 | 35.66 | 44.95 | 60.02 | 66.35 | 38.75 | 38.18 | 53.89 | 59.43 |
| PISCO† | 16× | 55.90 | 70.99 | 87.13 | 86.14 | 41.20 | 49.98 | 65.61 | 74.58 | 51.42 | 50.20 | 60.25 | 66.38 |
| DEX-Comp | 16× | **63.27** | **75.61** | **91.84** | **90.38** | **50.36** | **59.30** | **73.84** | **80.27** | **59.87** | **58.84** | **67.83** | **72.88** |
| **Top-30** | | | | | | | | | | | | | |
| Full context | - | 57.42 | 72.75 | 88.98 | 88.20 | 45.45 | 54.64 | 67.41 | 76.79 | 55.30 | 55.04 | 62.91 | 69.49 |
| LLMLingua-2 | 4× | 49.35 | 63.98 | 83.42 | 81.50 | 34.64 | 43.89 | 60.55 | 67.41 | 37.74 | 31.28 | 53.14 | 57.61 |
| PISCO† | 16× | 51.99 | 67.57 | 86.69 | 86.59 | 40.46 | 50.20 | 64.24 | 72.68 | 46.12 | 46.89 | 57.90 | 64.79 |
| DEX-Comp | 16× | **60.31** | **75.57** | **91.62** | **90.12** | **48.16** | **58.48** | **73.84** | **80.38** | **56.23** | **55.53** | **66.03** | **72.02** |

Table 2: Inference efficiency: TTFT, computation (GFLOPs), and peak GPU memory (MB).

| Top-$k$ | TTFT (ms) | | | GFLOPs | | | Memory (MB) | | |
|---|---|---|---|---|---|---|---|---|---|
| | RAG | DEX-Comp | Speedup | RAG | DEX-Comp | Reduction | RAG | DEX-Comp | Reduction |
| Top-5 | 373 | 84 | **4.42**× | 82,850 | 12,423 | **6.67**× | 16,100 | 14,242 | **1.13**× |
| Top-15 | 1,626 | 131 | **12.33**× | 245,650 | 22,654 | **10.84**× | 26,887 | 14,457 | **1.86**× |
| Top-30 | 4,782 | 201 | **23.73**× | 518,767 | 38,129 | **13.61**× | 59,328 | 14,766 | **4.02**× |

confirming the benefit of excluding teacher-failed examples from distillation. Applying HE to teacher-failed examples further improves accuracy to $72.02$ after PD, compared with $66.41$ when initialized from standard distillation, suggesting that PD provides a stronger initialization for exploration. Importantly, starting from the same PD checkpoint, exploration on teacher-failed examples achieves the highest accuracy ($72.02$), outperforming exploration on teacher-correct examples ($70.60$) or their mixture ($71.47$). Together, these results support the central design of DEX-Comp: teacher correctness is not merely a filtering criterion for distillation, but provides a useful signal for assigning imitation and outcome optimization to different training examples. Both stages also improve Resilience and Boost, especially Boost (Section 6.3), indicating more effective retrieval use; Figure 1(b,c) provides complementary diagnostics on teacher-failed questions.

### 6.2 Context Compression for Efficient RL

We compare DEX-Comp with untuned full-context RAG and full-context RL using matched RL data and procedures on eight GPUs (Table 4). At top-5, compression supports a $4\times$ larger maximum per-GPU batch ($64$ vs. $16$) and uses $65.8\%$ fewer GPU-hours ($13$, including PD, vs. $38$). It improves average accuracy over the untuned full-context model from $70.53$ to $72.54$, while full-context RL reaches $76.78$. Thus, compression reduces RL cost and enables training at top-30, where full-context RL exceeds our memory budget, while improving accuracy over untuned RAG ($72.02\%$ vs. $69.49\%$).

### 6.3 RAG Behavior Analysis

Resilience Rate measures correctness with and without retrieval, while Boost Rate measures retrieval-based corrections to initially incorrect answers (Cheng et al., 2024). DEX-Comp surpasses uncom-

Table 3: Top-30 ablation. Dataset scores and Acc are LLM-judged accuracy (%). RL subsets contain teacher-correct (Success), teacher-failed (Failure), or both (Mixed) questions.

| Initialization | RL subset | NQ | TriviaQA | HotpotQA | ASQA | PopQA | Average | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Acc | Resilience | Boost |
| Standard distill. | None | 67.57 | 86.59 | 50.20 | 72.68 | 46.89 | 64.79 | 85.25 | 45.96 |
| PD | None (w/o HE) | 70.96 | 88.74 | 55.32 | 74.79 | 48.95 | 67.75 | 87.63 | 49.56 |
| Standard distill. | Failure (HE) | 69.46 | 86.83 | 51.85 | 74.22 | 49.67 | 66.41 | 87.01 | 48.16 |
| PD | Success | 74.41 | 90.11 | 57.30 | 77.43 | 53.74 | 70.60 | 89.58 | 53.31 |
| PD | Mixed | 75.15 | 90.10 | 58.13 | 78.90 | 55.07 | 71.47 | 90.13 | 54.49 |
| **PD (DEX-Comp)** | **Failure (HE)** | **75.57** | **90.12** | **58.48** | **80.38** | **55.53** | **72.02** | **90.78** | **55.00** |

Table 4: Top-5 LLM-judged accuracy (%) and training cost on eight GPUs. Max. RL batch is the largest per-GPU batch fitting in memory; GPU-hours count total GPU training time.

| Configuration | NQ | TriviaQA | HotpotQA | ASQA | PopQA | Avg. | GPU-hours | Max. RL batch |
|---|---|---|---|---|---|---|---|---|
| Full context | 73.67 | 89.04 | 55.57 | 77.00 | 57.36 | 70.53 | – | – |
| Full context + RL | 80.23 | 92.36 | 65.00 | 82.28 | 64.02 | 76.78 | 38 | 16 |
| PD (16×) | 71.04 | 88.25 | 54.72 | 74.18 | 51.41 | 67.92 | 1.5 | – |
| PD + HE (16×) | 76.52 | 89.97 | 57.55 | 79.96 | 58.70 | 72.54 | 13 | 64 |

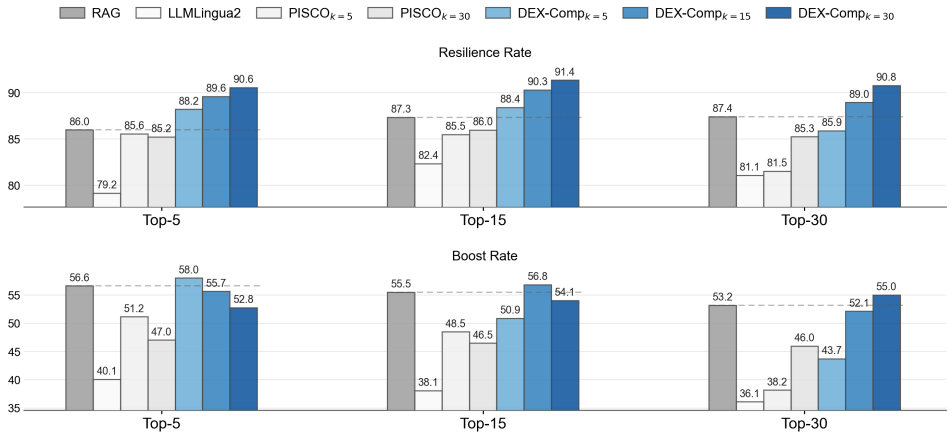



Figure 3: Resilience Rate (top) and Boost Rate (bottom) across retrieval depths. Subscripts denote the training top-$k$.

pressed RAG on both metrics at matched training and inference depths (Figure 3). Gains are larger for Resilience Rate, suggesting that compression acts as an implicit denoising mechanism.

### 6.4 Sensitivity to Compression and Retrieval Settings

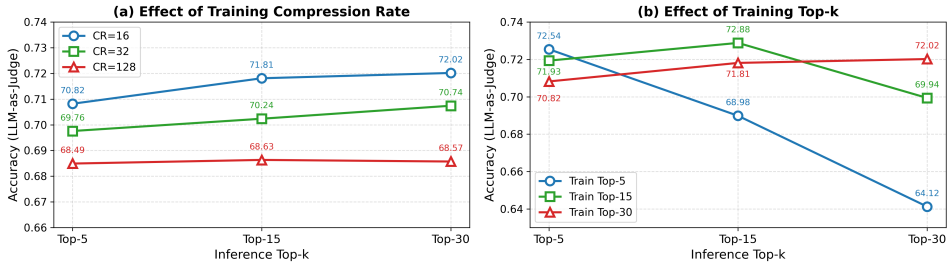



Figure 4: Effects of (a) compression rate and (b) training depth, with the other fixed.

Lower compression rates improve accuracy, with CR = 16 best across tested retrieval depths (Figure 4). Matching training and inference depths performs best; training at larger $k$ transfers more reliably to smaller depths than the reverse, favoring larger $k$ when deployment depth varies.

### 6.5 Generalization across Backbones and Datasets

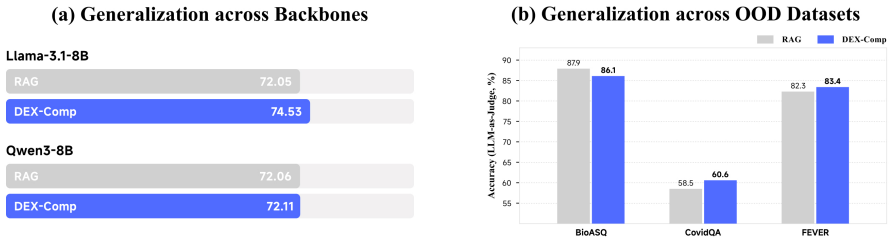



Figure 5: Generalization across (a) backbone families and (b) out-of-distribution datasets.

Figure 5(a) shows an improvement over the corresponding uncompressed RAG reference with Llama-3.1-8B and Qwen3-8B. For out-of-distribution evaluation, DEX-Comp performs better on CovidQA (Möller et al., 2020) and FEVER (Thorne et al., 2018), but lower on BioASQ (Krithara et al., 2023) (Figure 5(b)).

### 6.6 Analysis of Compression Embeddings

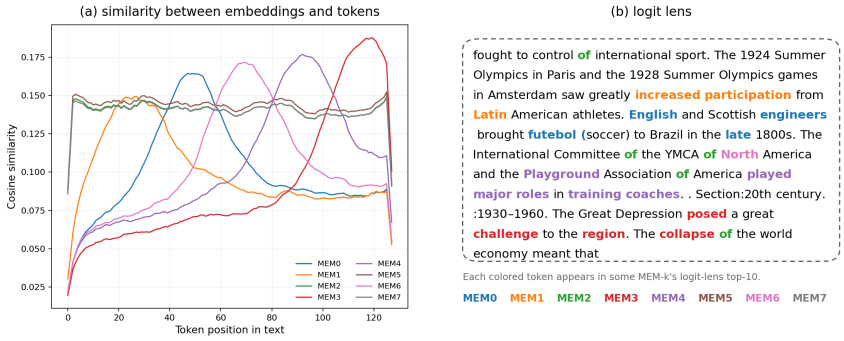



Figure 6: Compression embeddings: (a) token cosine similarity; (b) top logit-lens tokens.

We analyze learned embeddings using cosine similarity and logit lens projections (nostalgebraist, 2020). Embeddings exhibit spatial specialization, with some focusing on specific document regions and others capturing global information, suggesting complementary roles. Compared with PISCO trained by standard distillation, DEX-Comp exhibits markedly different embedding representations. The comparison is detailed in Appendix D.

## 7 Conclusion

We propose DEX-Comp, a two-stage training recipe for soft context compression that combines Pure Distillation with Hard Exploration. The recipe improves on the evaluated compression baselines without separate pretraining. At top-30 retrieval, DEX-Comp compresses the retrieved context by $16\times$ and accelerates inference by over $20\times$, while achieving performance comparable to or exceeding uncompressed RAG across five benchmarks. Further analyses confirm the contribution of each stage and demonstrate that DEX-Comp generalizes across backbone models and data distributions.

We believe this work provides a step toward understanding and improving soft context compression, and toward narrowing the gap between compressed and uncompressed RAG systems.

## AI USE STATEMENT

Our experimental use of LLMs includes teacher-response generation and Gemini 3 Flash judging for training-set partitioning and evaluation, as described in Sections 3–4 and Appendix A. Beyond these experimental uses, generative AI assistance was limited to sentence-level language editing, initial figure drafts, and experimental code. The authors take full responsibility for the final content of this paper, including its text, figures, code, and reported results.

## REFERENCES

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Aydar Bulatov, Yuri Kuratov, and Mikhail Burtsev. Recurrent memory transformer. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/47e288629a6996a17ce50b90a056a0e1-Abstract-Conference.html.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S. Abdelfattah, and Kai-Chiang Wu. Palu: KV-Cache compression with low-rank projection. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LWMS4pk2vK.

Shaoshen Chen, Yangning Li, Zishan Xu, Yongqin Zeng, Shunlong Wu, Xinshuo Hu, Zifei Shan, Xin Su, Jiwei Tang, Yinghui Li, and Hai-Tao Zheng. DAST: context-aware compression in LLMs via dynamic allocation of soft tokens. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pp. 20544–20552. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.1055/.

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 Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/c5cf13bfd3762821ef7607e63ee90075-Abstract-Conference.html.

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, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 3829–3846. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.232. URL https://doi.org/10.18653/v1/2023.emnlp-main.232.

Nadezhda Chirkova, Thibault Formal, Vassilina Nikoulina, and Stéphane Clinchant. Provence: efficient and robust context pruning for retrieval-augmented generation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TDy5Ih78b4.

Yu-Neng Chuang, Tianwei Xing, Chia-Yuan Chang, Zirui Liu, Xun Chen, and Xia Ben Hu. Learning to compress prompt in natural language formats. In Kevin Duh, Helena Gómez-Adorno, 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 7756–7767. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.429. URL https://doi.org/10.18653/v1/2024.naacl-long.429.

Tsz Ting Chung, Leyang Cui, Lemao Liu, Xinting Huang, Shuming Shi, and Dit-Yan Yeung. Selection-p: Self-supervised task-agnostic prompt compression for faithfulness and transferability. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, volume EMNLP 2024 of *Findings of ACL*, pp. 11057–11070. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-EMNLP.646. URL https://doi.org/10.18653/v1/2024.findings-emnlp.646.

Ziqiang Cui, Yunpeng Weng, Xing Tang, Peiyang Liu, Shiwei Li, Bowei He, Jiamin Chen, Xiuqiang He, and Chen Ma. CORE: lossless compression for retrieval-augmented LLMs via reinforcement learning. *CoRR*, abs/2508.19282, 2025. doi: 10.48550/ARXIV.2508.19282. URL https://doi.org/10.48550/arXiv.2508.19282.

Yuhong Dai, Jianxun Lian, Yitian Huang, Wei Zhang, Mingyang Zhou, Mingqi Wu, Xing Xie, and Hao Liao. Pretraining context compressor for large language models with embedding-based memory. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 28715–28732. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1394/.

Yixiong Fang, Tianran Sun, Yuling Shi, and Xiaodong Gu. AttentionRAG: Attention-guided context pruning in retrieval-augmented generation. *CoRR*, abs/2503.10720, 2025. doi: 10.48550/ARXIV.2503.10720. URL https://doi.org/10.48550/arXiv.2503.10720.

Weizhi Fei, Xueyan Niu, Guoqing Xie, Yingqing Liu, Bo Bai, and Wei Han. Efficient prompt compression with evaluator heads for long-context transformer inference. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/e356ed5f27885c79c7cb597bb1107c94-Abstract-Conference.html.

Yair Feldman and Yoav Artzi. Simple context compression: Mean-pooling and multi-ratio training. *CoRR*, abs/2510.20797, 2025. doi: 10.48550/ARXIV.2510.20797. URL https://doi.org/10.48550/arXiv.2510.20797.

Lang Feng, Fuchao Yang, Feng Chen, Xin Cheng, Haiyang Xu, Zhenglin Wan, Ming Yan, and Bo An. Agentocr: Reimagining agent history via optical self-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), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 5067–5086. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.230. URL https://doi.org/10.18653/v1/2026.acl-long.230.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023. doi: 10.48550/ARXIV.2312.10997. URL https://doi.org/10.48550/arXiv.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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=uREj4ZuGJE.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. Deeprag: Thinking to retrieve step by step for large language models. In *International Conference on Learning Representations*, volume 2026, pp. 117056–117071, 2026.

Shuyu Guo, Shuo Zhang, and Zhaochun Ren. Enhancing RAG efficiency with adaptive context compression. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 24061–24076. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.1307/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3929–3938. PMLR, 2020. URL http://proceedings.mlr.press/v119/guu20a.html.

Jitai Hao, Yuke Zhu, Tian Wang, Jun Yu, Xin Xin, Bo Zheng, Zhaochun Ren, and Sheng Guo. OmniKV: Dynamic context selection for efficient long-context LLMs. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=ulCAPXYXfa.

Jitai Hao, Qiang Huang, Yaowei Wang, Min Zhang, and Jun Yu. DeltaKV: Residual-based KV cache compression via long-range similarity. *CoRR*, abs/2602.08005, 2026. doi: 10.48550/ARXIV.2602.08005. URL https://doi.org/10.48550/arXiv.2602.08005.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

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, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jinwu Hu, Wei Zhang, Yufeng Wang, Yu Hu, Bin Xiao, Mingkui Tan, and Qing Du. Dynamic compressing prompts for efficient inference of large language models. *CoRR*, abs/2504.11004, 2025. doi: 10.48550/ARXIV.2504.11004. URL https://doi.org/10.48550/arXiv.2504.11004.

Taeho Hwang, Sukmin Cho, Soyeong Jeong, Hoyun Song, SeungYoon Han, and Jong C. Park. EXIT: context-aware extractive compression for enhancing retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pp. 4895–4924. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.253/.

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. *CoRR*, abs/2310.06825, 2023a. doi: 10.48550/ARXIV.2310.06825. URL https://doi.org/10.48550/arXiv.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, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 13358–13376. Association for

Computational Linguistics, 2023b. doi: 10.18653/V1/2023.EMNLP-MAIN.825. URL https://doi.org/10.18653/v1/2023.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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024.

Jiajie Jin, Xiaoxi Li, Guanting Dong, Yuyao Zhang, Yutao Zhu, Yongkang Wu, Zhonghua Li, Ye Qi, and Zhicheng Dou. Hierarchical document refinement for long-context retrieval-augmented generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 3502–3520. Association for Computational Linguistics, 2025a. URL https://aclanthology.org/2025.acl-long.176/.

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. *CoRR*, abs/2507.05633, 2025b. doi: 10.48550/ARXIV.2507.05633. URL https://doi.org/10.48550/arXiv.2507.05633.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.550. URL https://doi.org/10.18653/v1/2020.emnlp-main.550.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. BioASQ-QA: A manually curated corpus for biomedical question answering. *Scientific data*, 10(1):170, 2023.

Yuri Kuratov, Mikhail Arkhipov, Aydar Bulatov, and Mikhail Burtsev. Cramming 1568 tokens into a single vector and back again: Exploring the limits of embedding space capacity. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 19323–19339. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.948/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. *Trans. Assoc. Comput. Linguistics*, 7:452–466, 2019. doi: 10.1162/TACL_A_00276. URL https://doi.org/10.1162/tacl_a_00276.

Weronika Lajewska, Momchil Hardalov, Laura Aina, Neha Anna John, Hang Su, and Lluís Màrquez. Understanding and improving information preservation in prompt compression for llms. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 17520–17541. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.949. URL https://doi.org/10.18653/v1/2025.findings-emnlp.949.

Carlos Lassance, Hervé Déjean, Thibault Formal, and Stéphane Clinchant. Splade-v3: New baselines for SPLADE. *CoRR*, abs/2403.06789, 2024. doi: 10.48550/ARXIV.2403.06789. URL https://doi.org/10.48550/arXiv.2403.06789.

Yucheng Li. Unlocking context constraints of LLMs: Enhancing context efficiency of LLMs with self-information-based content filtering. *CoRR*, abs/2304.12102, 2023. doi: 10.48550/ARXIV.2304.12102. URL https://doi.org/10.48550/arXiv.2304.12102.

Yucheng Li, Bo Dong, Frank Guerin, and Chenghua Lin. Compressing context to enhance inference efficiency 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, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 6342–6353. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.391. URL https://doi.org/10.18653/v1/2023.emnlp-main.391.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/28ab418242603e0f7323e54185d19bde-Abstract-Conference.html.

Zongqian Li, Yinhong Liu, Yixuan Su, and Nigel Collier. Prompt compression for large language models: A survey. *CoRR*, abs/2410.12388, 2024b. doi: 10.48550/ARXIV.2410.12388. URL https://doi.org/10.48550/arXiv.2410.12388.

Zongqian Li, Yixuan Su, and Nigel Collier. 500xcompressor: Generalized prompt compression for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 25081–25091. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.1219. URL https://doi.org/10.18653/v1/2025.acl-long.1219.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. RA-DIT: retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=22OTbutug9.

Barys Liskavets, Maxim Ushakov, Shuvendu Roy, Mark Klibanov, Ali Etemad, and Shane K. Luke. Prompt compression with context-aware sentence encoding for fast and improved LLM inference. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 24595–24604. AAAI Press, 2025. doi: 10.1609/AAAI.V39I23.34639. URL https://doi.org/10.1609/aaai.v39i23.34639.

Guojie Liu, Yiqi Wang, Yanfeng Yang, Wenqi Fan, Songlei Jian, Jianfeng Zhang, and Jie Yu. Cognitive chunking for soft prompts: Accelerating compressor learning via block-wise causal masking. *CoRR*, abs/2602.13980, 2026a. doi: 10.48550/ARXIV.2602.13980. URL https://doi.org/10.48550/arXiv.2602.13980.

Quancai Liu, Haihui Fan, Jinchao Zhang, Xiangfang Li, Chuanrong Li, and Bo Li. Discomp: A two-stage prompt optimization framework combining task-agnostic and task-aware compression. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, Findings of ACL, pp. 1033–1044. Association for Computational Linguistics, 2025a. doi: 10.18653/V1/2025.FINDINGS-NAACL.58. URL https://doi.org/10.18653/v1/2025.findings-naacl.58.

Xin Liu, Runsong Zhao, Pengcheng Huang, Xinyu Liu, Junyi Xiao, Chunyang Xiao, Tong Xiao, Shengxiang Gao, Zhengtao Yu, and Jingbo Zhu. Autoencoding-free context compression for LLMs via contextual semantic anchors. *CoRR*, abs/2510.08907, 2025b. doi: 10.48550/ARXIV.2510.08907. URL https://doi.org/10.48550/arXiv.2510.08907.

Yunhao Liu, Zian Jia, Xinyu Gao, Kanjun Xu, and Yun Xiong. Rethinking soft compression in retrieval-augmented generation: A query-conditioned selector perspective. In Hakim Hacid, Yoelle Maarek, Francesco Bonchi, Ido Guy, and Emine Yilmaz (eds.), *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pp. 2441–2452. ACM, 2026b. 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, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pp. 15506–15521. Association for Computational Linguistics, 2025a. URL https://aclanthology.org/2025.findings-acl.800/.

Maxime Louis, Thibault Formal, Hervé Déjean, and Stéphane Clinchant. OSCAR: online soft compression and reranking. *CoRR*, abs/2504.07109, 2025b. doi: 10.48550/ARXIV.2504.07109. URL https://doi.org/10.48550/arXiv.2504.07109.

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 Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 9802–9822. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.546. URL https://doi.org/10.18653/v1/2023.acl-long.546.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. Nonparametric masked language modeling. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 2097–2118. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.132. URL https://doi.org/10.18653/v1/2023.findings-acl.132.

Youssef Mohamed, Mohamed Elhoseiny, Thibault Formal, and Nadezhda Chirkova. Xprovence: Zero-cost multilingual context pruning for retrieval-augmented generation. In Ricardo Campos, Adam Jatowt, Yanyan Lan, Mohammad Aliannejadi, Christine Bauer, Sean MacAvaney, Avishek Anand, Zhaochun Ren, Suzan Verberne, Nan Bai, and Masoud Mansoury (eds.), *Advances in Information Retrieval - 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 - April 2, 2026, Proceedings, Part II*, volume 16484 of *Lecture Notes in Computer Science*, pp. 470–480. Springer, 2026. doi: 10.1007/978-3-032-21300-6_38. URL https://doi.org/10.1007/978-3-032-21300-6_38.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, 2020.

Jesse Mu, Xiang Li, and Noah D. Goodman. Learning to compress prompts with gist tokens. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/3d77c6dcc7f143aa2154e7f4d5e22d68-Abstract-Conference.html.

nostalgebraist. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, Aug 2020. LessWrong blog post. Accessed: 2026-05-02.

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, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pp. 963–981. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.57. URL https://doi.org/10.18653/v1/2024.findings-acl.57.

Guanghui Qin, Corby Rosset, Ethan C. Chau, Nikhil Rao, and Benjamin Van Durme. Dodo: Dynamic contextual compression for decoder-only LMs. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 9961–9975. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.536. URL https://doi.org/10.18653/v1/2024.acl-long.536.

David Rau, Hervé Déjean, Nadezhda Chirkova, Thibault Formal, Shuai Wang, Stéphane Clinchant, and Vassilina Nikoulina. BERGEN: A benchmarking library for retrieval-augmented generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, Findings of ACL, pp. 7640–7663. Association for Computational Linguistics, 2024a. doi: 10.18653/V1/2024.FINDINGS-EMNLP.449. URL https://doi.org/10.18653/v1/2024.findings-emnlp.449.

David Rau, Shuai Wang, Hervé Déjean, and Stéphane Clinchant. Context embeddings for efficient answer generation in RAG. *CoRR*, abs/2407.09252, 2024b. doi: 10.48550/ARXIV.2407.09252. URL https://doi.org/10.48550/arXiv.2407.09252.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: retrieval-augmented black-box language models. In Kevin Duh, Helena Gómez-Adorno, 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), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pp. 8371–8384. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.NAACL-LONG.463. URL https://doi.org/10.18653/v1/2024.naacl-long.463.

Zhenmei Shi, Yifei Ming, Xuan-Phi Nguyen, Yingyu Liang, and Shafiq Joty. Discovering the gems in early layers: Accelerating long-context llms with 1000x input token reduction. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics, ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 13839–13857. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.FINDINGS-ACL.677. URL https://doi.org/10.18653/v1/2026.findings-acl.677.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: factoid questions meet long-form answers. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 8273–8288. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.566. URL https://doi.org/10.18653/v1/2022.emnlp-main.566.

Sijun Tan, Xiuyu Li, Shishir G. Patil, Ziyang Wu, Tianjun Zhang, Kurt Keutzer, Joseph Gonzalez, and Raluca A. Popa. LLoCO: Learning long contexts offline. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 17605–17621. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.975. URL https://doi.org/10.18653/v1/2024.emnlp-main.975.

Jiwei Tang, Jin Xu, Tingwei Lu, Zhicheng Zhang, Yiming Zhao, Lin Hai, and Hai-Tao Zheng. Perception compressor: A training-free prompt compression framework in long context scenarios. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pp. 4093–4108. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.229. URL https://doi.org/10.18653/v1/2025.findings-naacl.229.

Jiwei Tang, Zhijing Huang, Xinyu Zhang, Chen Jason Zhang, Jianxing Yu, Libin Zheng, Rui Meng, and Jian Yin. Beyond position bias: Shifting context compression from position-driven to semantic-driven. *CoRR*, abs/2605.09463, 2026a. doi: 10.48550/ARXIV.2605.09463. URL https://doi.org/10.48550/arXiv.2605.09463.

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Qingsong Lv, Runsong Zhao, Tingwei Lu, Langming Liu, Haibin Chen, Yujin Yuan, Hai-Tao Zheng, Wenbo Su, and Bo Zheng. Read as human: Compressing context via parallelizable close reading and skimming. 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), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 28393–28406. Association for Computational Linguistics, 2026b. doi: 10.18653/V1/2026.ACL-LONG.1309. URL https://doi.org/10.18653/v1/2026.acl-long.1309.

Jiwei Tang, Shilei Liu, Zhicheng Zhang, Yujin Yuan, Libin Zheng, Bo Zheng, et al. Comi: Coarse-to-fine context compression via marginal information gain. In *International Conference on Learning Representations*, volume 2026, pp. 46648–46666, 2026c.

Jiwei Tang, Zhicheng Zhang, Shunlong Wu, Jingheng Ye, Lichen Bai, Zitai Wang, Tingwei Lu, Lin Hai, Yiming Zhao, Hai-Tao Zheng, et al. Gmsa: Enhancing context compression via group merging and layer semantic alignment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28690–28704, 2026d.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.

Yujan Ting, Xu Tang, Terrence Chen, and Weijing Huang. Bridging the memorization-utilization gap: Near-lossless context compression via 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), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 14949–14972. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.682. URL https://doi.org/10.18653/v1/2026.acl-long.682.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. Learning to filter context for retrieval-augmented generation. *CoRR*, abs/2311.08377, 2023. doi: 10.48550/ARXIV.2311.08377. URL https://doi.org/10.48550/arXiv.2311.08377.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. In *International Conference on Learning Representations*, volume 2025, pp. 82731–82754, 2025.

David Wingate, Mohammad Shoeybi, and Taylor Sorensen. Prompt compression and contrastive conditioning for controllability and toxicity reduction in language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5621–5634. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.412. URL https://doi.org/10.18653/v1/2022.findings-emnlp.412.

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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=mlJLVigNHp.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. 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, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Howard Yen, Tianyu Gao, and Danqi Chen. Long-context language modeling with parallel context encoding. 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), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 2588–2610. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.142. URL https://doi.org/10.18653/v1/2024.acl-long.142.

Chanwoong Yoon, Taewhoo Lee, Hyeon Hwang, Minbyul Jeong, and Jaewoo Kang. Compact: Compressing retrieved documents actively for question answering. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 21424–21439. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.1194. URL https://doi.org/10.18653/v1/2024.emnlp-main.1194.

JungMin Yun and YoungBin Kim. Itercomp: Reasoning-aware adaptive prompt compression for multi-hop question answering. 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), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 33827–33840. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1559. URL https://doi.org/10.18653/v1/2026.acl-long.1559.

Peitian Zhang, Zheng Liu, Shitao Xiao, Ninglu Shao, Qiwei Ye, and Zhicheng Dou. Long context compression with activation beacon. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=1eQT9OzfNQ.

Runsong Zhao, Xin Liu, Xinyu Liu, Pengcheng Huang, Chunyang Xiao, Tong Xiao, and JingBo Zhu. Position ids matter: An enhanced position layout for efficient context compression in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 17715–17734. Association for Computational Linguistics, 2025a. URL https://aclanthology.org/2025.findings-emnlp.962/.

Runsong Zhao, Shilei Liu, Jiwei Tang, Langming Liu, Haibin Chen, Weidong Zhang, Yujin Yuan, Tong Xiao, JingBo Zhu, Wenbo Su, and Bo Zheng. Comet: Collaborative memory transformer for efficient long context modeling. 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), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 38384–38402. Association for Computational Linguistics, 2026a. doi: 10.18653/V1/2026.ACL-LONG.1779. URL https://doi.org/10.18653/v1/2026.acl-long.1779.

Runsong Zhao, Shilei Liu, Jiwei Tang, Langming Liu, Haibin Chen, Weidong Zhang, Yujin Yuan, Tong Xiao, Jingbo Zhu, Wenbo Su, and Bo Zheng. Poc: Performance-oriented context compression for large language models via performance prediction. *CoRR*, abs/2603.19733, 2026b. doi: 10.48550/ARXIV.2603.19733. URL https://doi.org/10.48550/arXiv.2603.19733.

Yi Zhao, Zuchao Li, Hai Zhao, Baoyuan Qi, and Guoming Liu. DAC: A dynamic attention-aware approach for task-agnostic prompt compression. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 19395–19407. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.acl-long.952/.

Yunlong Zhao, Haoran Wu, and Bo Xu. Leveraging attention to effectively compress prompts for long-context llms. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference*

*on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 26048–26056. AAAI Press, 2025c. doi: 10.1609/AAAI.V39I24.34800. URL https://doi.org/10.1609/aaai.v39i24.34800.

## A LLM-AS-JUDGE PROMPT

The following prompt is used to query the judge model for the LLM-as-judge accuracy metric reported in our experiments.

```
You are an evaluation tool. Just answer by {{ Yes }} or {{ No }}.

Given a question, a list of golden answers (any one of them is
considered correct), and an AI-generated answer, judge whether the AI-
generated answer is correct - it is correct if it matches ANY ONE of
the golden answers.

Question: {question}
Golden answers: {ground_truth}

Generated answer: {model_answer}
Response: {{
```

## B PURE DISTILLATION HYPERPARAMETERS

Pure Distillation (§3.2) is performed on the teacher-correct subset $\mathcal{S}^+$, where correctness is determined by Gemini 3 Flash as described in Section 4.3. The core hyperparameters are listed in Table 5.

Table 5: Hyperparameters for Pure Distillation.

| **Hyperparameter** | **Assignment** |
|---|---|
| optimizer | AdamW |
| learning rate | $1\times10^{-4}$ |
| warmup ratio | 0.05 |
| weight decay | 0.1 |
| gradient clipping | 1.0 |
| epochs | 1 |
| per-device batch size | 2 |
| gradient accumulation steps | 1 |
| effective batch size | 16 |
| max response length | 128 |
| distributed setup | DeepSpeed (8 GPUs) |
| random seed | 42 |

## C HARD EXPLORATION HYPERPARAMETERS

Hard Exploration (§3.3) initializes from the PD checkpoint and runs GRPO on the teacher-failed subset $\mathcal{S}^-$. The core hyperparameters are listed in Table 6.

## D COMPARISON WITH STANDARD DISTILLATION

We compare the compression embeddings learned by PISCO through standard distillation (Figure 7) with those of DEX-Comp (Figure 6). Both visualizations use the same passage and show eight

Table 6: Hyperparameters for Hard Exploration.

| Hyperparameter | Assignment |
|---|---|
| optimizer | AdamW |
| learning rate | $5 \times 10^{-6}$ |
| warmup ratio | 0.05 |
| weight decay | 0.01 |
| gradient clipping | 1.0 |
| gradient checkpointing | True |
| epochs | 2 |
| training micro batch size | 8 |
| per-GPU policy batch size | 64 |
| gradient accumulation steps | 1 |
| GRPO group size $G$ | 16 |
| clipping range $\varepsilon$ | 0.2 |
| KL coefficient $\beta$ | 0.02 |
| reward function | binary CEM |
| rollout temperature | 1 |
| rollout top-$p$ | 0.95 |
| max generation length | 32 |
| random seed | 42 |

memory embeddings. We compare the distribution of information across each model's embeddings, whose indices are model-specific.

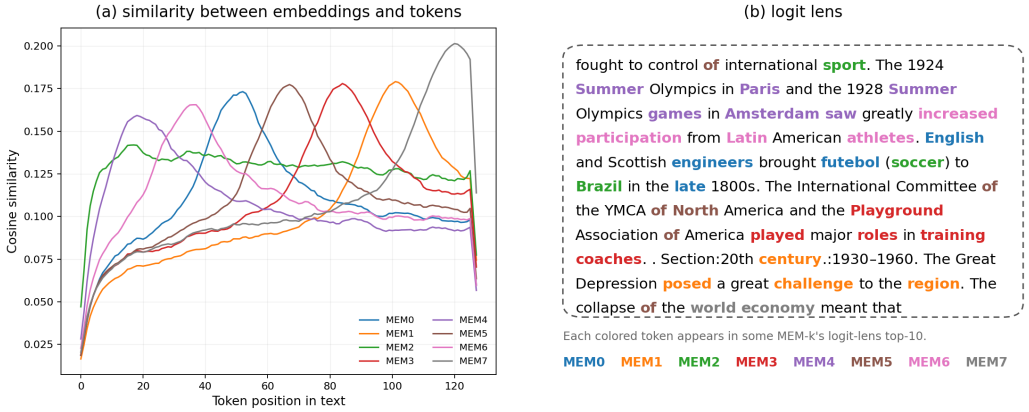



Figure 7: PISCO compression embeddings learned by standard distillation, for the same passage as Figure 6. (a) Cosine similarity between embeddings and document tokens. (b) Document tokens appearing among the top-10 logit-lens vocabulary projections of the memory embeddings. Colors identify embeddings within each model.

**Global and local information.** In this example, PISCO has one embedding with a broadly distributed similarity profile (MEM2), while the other seven exhibit localized peaks. DEX-Comp has three embeddings (MEM2, MEM5, and MEM7) whose similarities remain relatively high across most of the document tokens, alongside five embeddings with localized peaks. This pattern suggests that DEX-Comp allocates more embeddings to global information while retaining representations of specific text regions.

**Logit-lens projections.** The projections differ in both the highlighted vocabulary and its distribution across embeddings. PISCO highlights the Olympic place names *Paris* and *Amsterdam*, which do not appear in DEX-Comp's displayed overlay. DEX-Comp instead highlights *major* in the coach-training passage. Shared tokens are also distributed differently: *increased participation* and *Latin* appear under MEM6 in PISCO and MEM1 in DEX-Comp, while much of the coach-training passage appears

under MEM3 in PISCO and MEM4 in DEX-Comp. Together with the more broadly distributed similarity profiles, these differences reveal a distinct organization of compressed representations in this example.