Conflict-Aware Soft Prompting for Retrieval-Augmented Generation

Eunseong Choi, June Park, Hyeri Lee, Jongwuk Lee*
Sungkyunkwan University, Republic of Korea
{eunseong, pj00515, bluepig94, jongwuklee}@skku.edu

Abstract

Retrieval-augmented generation (RAG) enhances the capabilities of large language models (LLMs) by incorporating external knowledge into their input prompts. However, when the retrieved context contradicts the LLM's parametric knowledge, it often fails to resolve the conflict between incorrect external context and correct parametric knowledge, known as *context-memory conflict*. To tackle this problem, we introduce ***C**onflict-**A**ware **RE**trieval-Augmented Generation (**CARE**)*, consisting of a *context assessor* and a *base LLM*. The context assessor encodes external context into compact *memory embeddings*. Through *grounded/adversarial soft prompting*, the context assessor is trained to discern unreliable context and capture a guidance signal that directs reasoning toward the more reliable knowledge source. Extensive experiments show that CARE effectively mitigates context-memory conflicts, leading to an average performance gain of 5.0% on QA and fact-checking benchmarks, establishing a promising direction for trustworthy and adaptive RAG systems[1].

1 Introduction

Retrieval-augmented generation (RAG) serves as an effective strategy for enhancing large language models (LLMs) by grounding them in external information (Lewis et al., 2020; Gao et al., 2023; Huang and Huang, 2024). However, when the retrieved context contradicts the LLM's internal knowledge, it causes a critical vulnerability: the challenge of resolving the conflict between incorrect external context and correct internal knowledge, known as *context-memory conflict*.

This challenge is exacerbated when incorrect yet highly ranked contexts serve as hard negatives. Conventional RAG, *i.e.*, simply appending retrieved context to the prompt, struggles to discriminate between incorrect external context and correct parametric knowledge (Ren et al., 2025). This misalignment leads to overriding correct internal representations, resulting in substantial performance degradation on questions that the model initially answered correctly. As shown in Figure 1, we observed significant performance drops of 25.1-49.1% across state-of-the-art LLMs when negative contexts were added to questions for which the LLM had already generated the correct answer without external context.

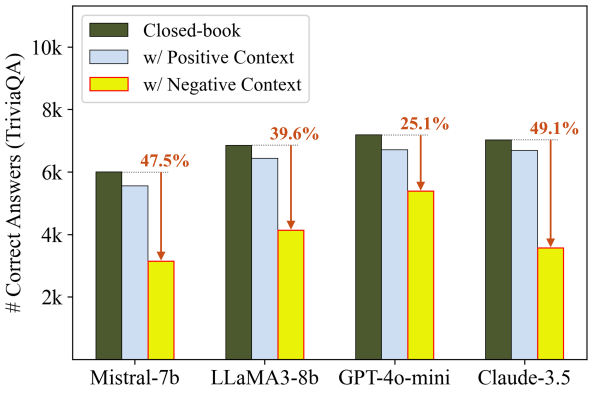

Figure 1: LLMs struggle to resolve context-memory conflict. Green bars show the number of questions correctly answered without retrieval in a closed-book setting. Blue and yellow bars show performance when provided with a positive or negative context, respectively.

To mitigate context-memory conflict, existing studies such as adaptive retrieval (Ren et al., 2025; Baek et al., 2025) and the decoding strategies (Zhao et al., 2024; Han et al., 2025) adjust the influence of external context either before or during answer generation. However, due to the LLM's limited capacity in detecting conflicts, it is susceptible to misleading contextual inputs that contradict the LLM's parametric knowledge. Recently, robust training has equipped LLMs, enabling them to identify conflicts (Asai et al., 2024; Wang et al., 2024). As shown in Figure 2(a), it enables the LLM to dis-

* Corresponding author
[1]https://github.com/eunseongc/CARE

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26981–26995
November 4-9, 2025

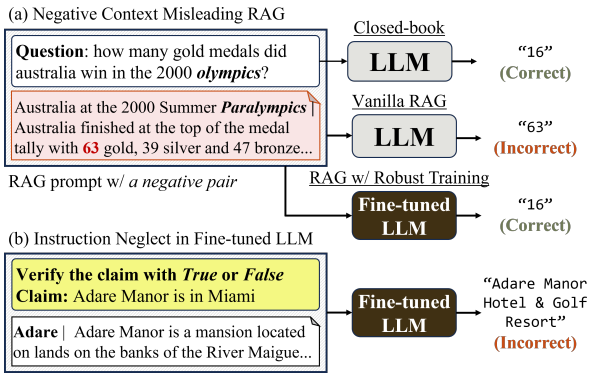

Figure 2: (a) Vanilla RAG fails with negative context despite having the correct information in parametric knowledge. (b) Robust training can cause the LLM to disregard instructions and suffer from catastrophic forgetting, particularly when transitioning between tasks like question answering and fact verification.

Approach	Soft Decision	Conflict Awareness	Generality
Adaptive Retrieval	✗	✗	✓
Decoding Strategy	✓	✗	✓
Robust Training	✓	✓	✗
CARE (Ours)	✓	✓	✓

Table 1: Comparisons for existing studies to resolve knowledge conflict. Our method achieves comprehensive coverage across all criteria.

cern conflicts and assess the confidence of external contexts, *i.e.*, whether to rely on them during generation. Although it demonstrates promising in-domain performance gains, it incurs the catastrophic forgetting (Luo et al., 2023; Yang et al., 2024) problem, which significantly impairs the generalization performance of the LLM. As shown in Figure 2(b), the robust fine-tuning of QA datasets results in the model forgetting knowledge related to tasks beyond the specific datasets.

This motivates our central research question: *How can we incorporate conflict awareness into LLMs while preserving their general capabilities?* To this end, we propose ***C**onflict-**A**ware **RE**trieval-Augmented Generation (**CARE**)*, comprising two components: a *context assessor* and a *base LLM*. The context assessor, instantiated from the base LLM itself, is designed to identify knowledge conflicts. Inspired by soft prompting (Ge et al., 2024), the context assessor encodes external context into compact, trainable memory tokens, referred to as *soft context embeddings*.

Specifically, the context assessor is trained using a conflict-aware training strategy. First, we perform reconstruction pre-training to enable the context assessor to encode *memory embeddings* from raw context. Next, we fine-tune the assessor to discern between accurate and conflicting knowledge via *grounded/adversarial soft prompting*. When the base LLM produces an incorrect answer relying solely on its parametric knowledge, we provide a positively retrieved passage to supervise the assessor. Conversely, when the LLM answers correctly without retrieval, we pair it with a hard negative context to discourage unnecessary reliance on retrieved context. This training strategy enables the assessor to guide the LLM along the appropriate reasoning path, balancing retrieved and parametric knowledge. Notably, CARE does not require any fine-tuning of the base LLM, preserving the general-purpose ability of the base LLM.

We summarize our key contributions as follows.

- We examine how the conflict between external context and parametric memory hinders the conventional RAG system.
- We propose CARE via grounded/adversarial soft prompting to learn context embeddings that encode both the context and its implicit confidence, thereby preventing the LLM from being misled by conflicting knowledge.
- Experimental results show that CARE significantly improves the robustness of RAG systems, increasing the overall performance by up to 5.0% over existing methods on QA and fact-checking benchmarks.

2 Related Work

2.1 Context-Memory Conflict

Recent studies have focused on improving the conflict between external context and internal knowledge. Specifically, they are categorized into three directions: (i) Adaptive Retrieval, (ii) Decoding Strategies, and (iii) Robust Training.

Adaptive Retrieval. It aims to selectively incorporate external context only when the LLM lacks sufficient knowledge, mitigating the conflict between retrieved and parametric knowledge. The decision is generally made through prompting the LLM to judge its uncertainty (Ren et al., 2025), estimating confidence from hidden states (Su et al., 2024; Yao et al., 2024), or using an external module trained to predict retrieval necessity (Wang et al., 2023; Jeong et al., 2024; Baek et al., 2025). However, it is inherently difficult for the LLM to accurately assess the boundaries of its knowledge in the process

of making discrete retrieval decisions (Xiong et al., 2024).

Decoding Strategy. A representative work, CAD (Shi et al., 2024) adjusts the model's output distribution by contrasting output probability distributions with and without the context. Subsequent studies use additional information to dynamically adjust the weights of the contrastive decoding distribution according to the strength of the knowledge conflict (Yuan et al., 2024; Han et al., 2025). Since these methods combine distributions that already reflect conflicting information, it is crucial to incorporate conflict-awareness within the model itself rather than relying solely on the decoding stage.

Robust Training. It trains the LLM to assess the reliability of retrieved documents based on its internal knowledge. A representative method performs adversarial training, where negative documents are introduced during fine-tuning to help the model recognize contradictions and assess context reliability (Yoran et al., 2024; Fang et al., 2024). Another line of work explicitly trains the LLM to acquire new capabilities, such as evaluating the relevance of retrieved documents and deciding whether retrieval is necessary (Wang et al., 2024; Asai et al., 2024). However, fine-tuning LLMs risks catastrophic forgetting, which can degrade their generalization performance and erode parametric knowledge. In contrast, our method addresses the knowledge conflict by leveraging conflict-aware context representations, thereby preserving the generality of LLMs. Table 1 presents a comparison between CARE and existing approaches across the three criteria.

2.2 Soft Prompting

Soft prompting aims to encode lengthy text into a few trainable continuous embeddings (Chevalier et al., 2023; Mu et al., 2023; Qin and Durme, 2023). Ge et al. (2024) introduces *memory slots* appended to the input context, leveraging their final hidden states to reconstruct the original context. Recently, Cheng et al. (2024) introduced a lightweight MLP to transform retrieval embeddings into document-level soft prompt vectors. While previous studies have primarily used soft prompts to compress retrieved contexts or instructions, we repurpose them to balance the influence of external information against internal knowledge. Specifically, we encode soft context embeddings that reflect the reliability of the retrieved content. Our approach offers a straightforward solution to the context-memory conflict by improving the context representation directly, rather than relying on an additional mechanism at the retrieval or decoding stage.

3 Proposed Method: CARE

In this section, we present ***C**onflict-**A**ware **RE**trieval-Augmented Generation (**CARE**)*, addressing knowledge conflicts in conventional RAG systems. As illustrated in Figure 3, CARE comprises two main components: a *context assessor* and a *base LLM*. The context assessor inherits its parametric knowledge from the base LLM, encoding *memory embeddings*, *i.e.*, soft representations of the input context, and plays a critical role in assessing the reliability of retrieved contexts.

Specifically, the training of the context assessor proceeds in two stages. For reconstruction pre-training, it learns to encode contextual information into compact memory embeddings (Section 3.1). For conflict-aware fine-tuning, it is optimized to identify knowledge conflicts and guide the base LLM toward the most trustworthy source of information (Section 3.2).

3.1 Reconstruction Pre-training

The goal of the pre-training stage is to train the context assessor to represent the retrieved context as memory token embeddings. We adopt reconstruction-based soft prompting methods (Ge et al., 2024; Cheng et al., 2024).

we append K learnable memory tokens, denoted as $M = \langle m_1, \ldots, m_K \rangle$, to the original input context $C = [c_1, \ldots, c_n]$. The resulting sequence serves as the input for pre-training.

$$X_{\text{PT}} = [c_1, c_2, \ldots, c_n, \langle m_1, \ldots, m_K \rangle]. \quad (1)$$

We feed an input token sequence into the context assessor, incorporating a LoRA adapter built upon the base LLM. The final hidden states corresponding to the memory tokens M are obtained from a single forward pass of the decoder, referred to as *memory embeddings* ($\mathbf{E}_{\text{mem}}$). These embeddings serve as compressed representations of the input context and are used as soft prompts for downstream text generation.

$$\mathbf{E}_{\text{mem}} = f_{\phi,\theta}(X_{\text{PT}})_{n+1:n+K} \in \mathbb{R}^{K \times d}, \quad (2)$$

where ϕ denotes frozen LLM parameters, and θ represents the learnable parameters of the LoRA adapters (Hu et al., 2022) and memory tokens.

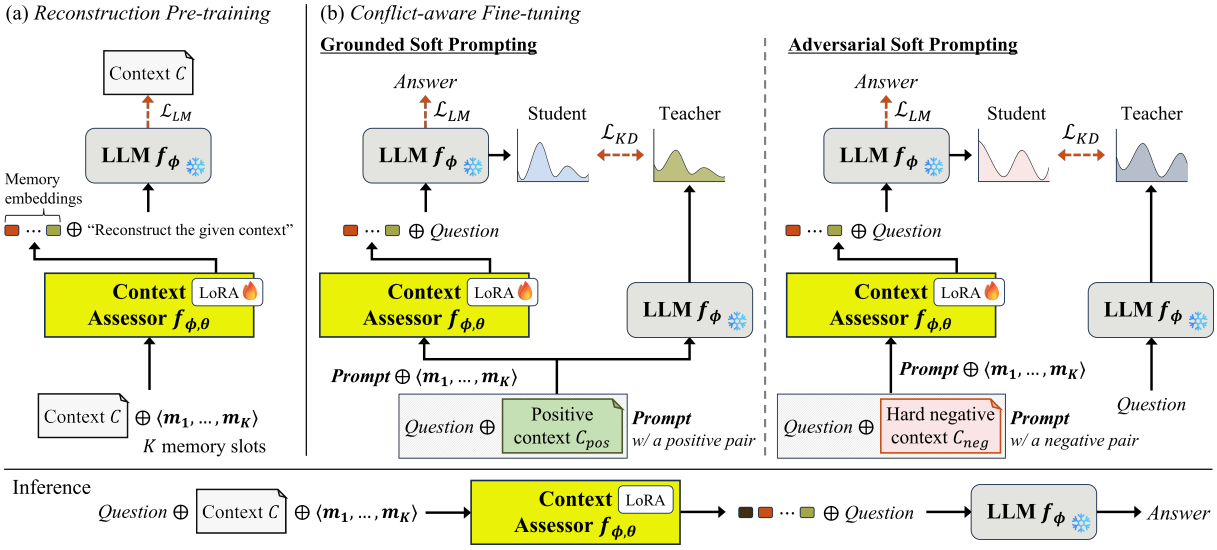

Figure 3: Overall framework of **C**onflict-**A**ware **RE**trieval-augmented generation (**CARE**). (a) The reconstruction pretraining learns to compress contextual information into memory embeddings. (b) In the conflict-aware fine-tuning stage, adversarial soft prompting exposes the Context Assessor to misleading contexts, encouraging the memory embeddings to reflect their reliability.

We train the parameters θ by minimizing a reconstruction loss computed using the frozen LLM parameters ϕ, which encourages $\mathbf{E}_{\text{mem}}$ to capture the accurate information required to reconstruct the input context C. The reconstruction loss is defined as the negative log-likelihood of generating each token conditioned on the memory embeddings $\mathbf{E}_{\text{mem}}$ and instruction I_{recon}[2]:

$$\mathcal{L}_{\text{PT}} = -\sum_{i=1}^{n} \log P_\phi(c_i \mid \mathbf{E}_{\text{mem}}, \text{I}_{\text{recon}}, c_{<i}). \quad (3)$$

3.2 Conflict-aware Fine-tuning

We fine-tune the context assessor to be sensitive to *context-memory conflicts*. The goal is to produce memory embeddings that not only compress the retrieved context but also reflect its reliability, *i.e.*, whether the LLM should rely on it. To achieve this, we employ an instruction-following dataset with question-answer pairs to expose the discrepancy between external and parametric knowledge.

The context assessor takes an input sequence consisting of question $Q = [q_1, \ldots, q_n]$, context $C = [c_1, \ldots, c_n]$, and learnable memory tokens $M = \langle m_1, \ldots, m_K \rangle$, to assess the context based on the question.

$$X_{\text{FT}} = [q_1, \ldots, q_m, c_1, \ldots, c_n, \langle m_1, \ldots, m_K \rangle]. \quad (4)$$

[2]We adopt the instruction set from Cheng et al. (2024) for reconstruction.

As in the pre-training stage, the context assessor extracts memory embeddings $\mathbf{E}_{\text{mem}} \in \mathbb{R}^{K \times d}$ by encoding the input sequence:

$$\mathbf{E}_{\text{mem}} = f_{\phi,\theta}(X_{\text{FT}})_{n+m+1:n+m+K}. \quad (5)$$

To train the context assessor to detect such conflicts, we simulate them using correctness signals from a closed-book setting: when the LLM fails to answer correctly without context, and when it succeeds. If the parametric knowledge is insufficient, we apply *grounded soft prompting* to guide the LLM toward the external context. When the LLM already contains the relevant knowledge, we apply *adversarial soft prompting* with a hard negative context and train the context assessor to encode it in a way that reduces the influence, allowing the LLM to favor internal knowledge.

Grounded Soft Prompting. It provides the context assessor with positive supervision signals about useful external knowledge. For questions that the LLM fails to answer in a closed-book setting, we pair them with a positive context C_{pos} that contains the answer span. In this setup, C_{pos} is treated as a reliable context source. The context assessor is trained to reflect this reliability in memory embeddings, allowing the model to recognize and represent helpful external knowledge, particularly when parametric knowledge is insufficient.

Adversarial Soft Prompting. We adopt supervision for identifying and down-weighting unreliable external information. For questions that the LLM

already answers correctly in a closed-book setting, we construct conflict scenarios by pairing the question with a hard negative passage C_{neg}, which is topically relevant but does not contain the correct answer. The context assessor is trained to reflect low reliability for such passages in its memory embeddings, effectively learning to recognize when external context contradicts parametric knowledge. As a result, misleading information has less influence on the generation, and the LLM relies more on its internal knowledge.

Training Objective. We optimize the context assessor parameters θ using two complementary objectives during fine-tuning: a *language modeling loss* and a *knowledge distillation loss*.

The language modeling (LM) loss $\mathcal{L}_{\text{LM}}$ ensures that the memory embeddings $\mathbf{E}_{\text{mem}}$ support accurate answer generation by the frozen LLM f_ϕ. Since both helpful and misleading contexts are used during fine-tuning, the memory embeddings must encode not only contextual information but also reliability, allowing the LLM to either utilize or disregard the context as needed. Given a target output $A = [a_1, \ldots, a_k]$, the LM loss is defined as follows.

$$\mathcal{L}_{\text{LM}} = -\sum_{i=1}^{k} \log P_\phi(a_i \mid \mathbf{E}_{\text{mem}}, Q, a_{<i}). \quad (6)$$

Although it encourages the context assessor to produce memory embeddings that support accurate responses, it does not supervise how these embeddings should be used. That is, solely relying on the LM loss cannot explicitly distinguish whether the model should rely on external information or its parametric knowledge.

To address this, we adopt a knowledge distillation (KD) loss using scenario-dependent supervision. Given a target output sequence A, the student distribution at i-th decoding step $P^{(i)}_{\text{student}}$ is computed using the frozen LLM ϕ, conditioned on the question Q and the memory embeddings $\mathbf{E}_{\text{mem}}$ as follows.

$$P^{(i)}_{\text{student}} := P_\phi(a_i \mid \mathbf{E}_{\text{mem}}, Q, a_{<i}). \quad (7)$$

The teacher distribution is defined depending on whether the given context is helpful or misleading:

$$P^{(i)}_{\text{teacher}} = \begin{cases} P_\phi(a_i \mid Q, C_{\text{pos}}, a_{<i}) & \text{(Ground.)} \\ P_\phi(a_i \mid Q, a_{<i}) & \text{(Advers.).} \end{cases} \quad (8)$$

We then adaptively minimize the KL divergence between these token-level distributions.

$$\mathcal{L}_{\text{KD}} = \sum_{i=1}^{k} \text{KL}\left(P^{(i)}_{\text{student}} \parallel P^{(i)}_{\text{teacher}}\right) \quad (9)$$

This scenario-specific supervision forms a dual distillation objective, training the memory embeddings to guide the LLM's reasoning based on the utility of the retrieved context. Finally, we obtain a fine-tuning objective as a weighted sum.

$$\mathcal{L}_{\text{FT}} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{KD}}. \quad (10)$$

4 Experiments Setup

Datasets. We conduct extensive evaluation across three distinct tasks: *Open-domain QA*, *Long-form QA*, and *Fact verification*. (i) For the Open-domain QA task, we selected Natural Questions (NQ) (Kwiatkowski et al., 2019), along with TriviaQA (Joshi et al., 2017) and WebQuestions (WebQA) (Berant et al., 2013). (ii) To assess long-form generation capabilities, we utilized TruthfulQA (Lin et al., 2022), which requires the model to produce truthful answers when faced with questions to elicit false or misleading responses. (iii) Lastly, we utilized the FactKG dataset (Kim et al., 2023), which requires the model to determine the factuality of claims by leveraging knowledge from either parametric or the external source. See Table 9 and Appendix A.2 for dataset statistics and prompts used for each task.

Baseline Methods. We compare CARE against several existing methods to address knowledge conflicts using different strategies.

- **Robust training**: It enhances the resilience of the LLMs through adversarial learning. For instance, **RetRobust** (Yoran et al., 2024) teaches the model to ignore irrelevant or distracting retrieved contexts, making LLM more robust. As a variant, we also apply our conflict-aware training strategy directly to the base LLM without using a context assessor, denoted as **Direct FT**.
- **Decoding-based strategy**: **CAD** (Shi et al., 2024) aims to assess and prioritize the reliability of retrieved documents during inference. **ADA-CAD** (Han et al., 2025) utilizes confidence-aware decoding to downweight unreliable contexts.
- **Adaptive retrieval**: **Adaptive-RAG** (Jeong et al., 2024) trains a classifier and question difficulty. **SKR-kNN** (Wang et al., 2023) embeds examples of correct and incorrect LLM generations

LLM	Method	Open-Domain QA (Span EM) NQ	TriviaQA	webQA	Long-form QA (F1 / ROUGE-L) TruthfulQA		Fact checking (Acc) FactKG	Average
Mistral-7B	*Fine-tuning LLMs*							
	Direct FT	0.469	0.708	0.389	0.102	0.069	0.542	0.380
	RetRobust	0.459	0.719	0.412	0.152	0.134	0.583	0.410
	Without Fine-tuning LLMs							
	Closed-book	0.290	0.577	0.366	0.273	0.249	0.531	0.381
	RAG	0.419	0.659	0.372	0.277	0.251	**0.639**	0.436
	CAD	0.402	0.631	0.352	0.243	0.216	0.598	0.407
	ADACAD	0.417	0.653	0.364	0.267	0.242	0.636	0.430
	Adaptive-RAG	0.402	0.631	0.367	0.275	0.252	0.633	0.427
	SKR-kNN	0.406	0.639	0.396	0.278	0.253	0.630	0.434
	Priori Judgment	0.422	0.682	0.378	0.274	0.251	0.600	0.435
	CARE (Ours)	**0.447**	**0.696**	**0.432**	**0.279**	**0.256**	0.638	**0.458**
LLaMA-3-8B	*Fine-tuning LLMs*							
	Direct FT	0.472	0.711	0.360	0.100	0.067	0.305	0.336
	RetRobust	0.461	0.726	0.380	0.081	0.061	0.644	0.392
	Without Fine-tuning LLMs							
	Closed-book	0.345	0.630	**0.465**	**0.266**	0.239	0.637	0.430
	RAG	0.447	0.684	0.377	0.247	0.225	0.661	0.440
	CAD	0.394	0.613	0.326	0.164	0.142	0.610	0.375
	ADACAD	0.428	0.666	0.362	0.214	0.193	0.667	0.422
	Adaptive-RAG	0.458	0.690	0.438	0.245	0.223	0.661	0.452
	SKR-kNN	0.449	0.675	0.427	0.246	0.224	0.660	0.447
	Priori Judgment	0.458	**0.704**	0.406	0.254	0.231	0.666	0.453
	CARE (Ours)	**0.465**	0.700	0.445	0.264	**0.243**	**0.686**	**0.467**

Table 2: Evaluation results on Open-Domain QA (NQ, TriviaQA, WebQA), Long-form QA (TruthfulQA), and Fact Checking (FactKG), using Mistral-7B-Instruct and LLaMA-3-8B-Instruct as base LLMs. The best performance is marked in **bold**, and the second-best is underlined, among retrieval-augmented strategies that do not directly fine-tune the LLMs, including Decoding Strategy and Adaptive Retrieval approaches.

and uses kNN search over these embeddings to decide whether retrieval is needed. **Priori Judgement** (Ren et al., 2025) examines the LLM's knowledge boundaries through prompts to determine whether to use external passages. At inference time, these methods decide whether to use external evidence and select either a closed-book or RAG response as the final answer.

Evaluation Metrics. We adopt task-specific evaluation metrics that align with each benchmark. For Open-domain QA datasets, we report Span EM, measuring whether any ground-truth answer appears within the generated output. Unlike standard exact match, this evaluates containment rather than strict equivalence, making it better suited for assessing the knowledge encoded in LLMs. For the long-form QA task, we use F1 and ROUGE-L scores. For the fact verification task *FactKG*, the model is required to generate either "true" or "false." We report accuracy as the evaluation metric.

Implementation Details. We mainly employ two base LLMs for CARE: *Mistral-7B* (Jiang et al., 2023)[3] and *LLaMA-3-8B* (Dubey et al., 2024)[4], both in their instruct versions. To further assess the generality, we also evaluate CARE on Qwen-based model (Yang et al., 2025), as detailed in Appendix B. We employ ColBERTv2 (Santhanam et al., 2022), to retrieve top-1 context from Wikipedia for external knowledge through all datasets. To train the context accessor module efficiently, we use the LoRA (Hu et al., 2022) adapter in pre-training and fine-tuning. For the pre-training stage, we adopted prompts from xRAG (Cheng et al., 2024) and used two million randomly sampled contexts from the December 2021 Wikipedia dump. We use the *NQ train* for fine-tuning and hold out 10% for validation. Since CARE requires explicit positive and negative contexts, we filter out questions for which none of the top 100 retrieved passages contain an answer span. We set the number of memory tokens K as 16 for all experiments. We use a zero-shot setup to precisely

[3] mistralai/Mistral-7B-Instruct-v0.2
[4] meta-llama/Meta-Llama-3-8B-Instruct

evaluate conflict, in which the model does not receive in-context examples. Refer to Appendix A.1 for detailed hyperparameters.

5 Results and Analysis

5.1 Main Results

Table 2 compares CARE with baselines across tasks using two recent LLMs. CARE achieves the best overall performance, outperforming standard RAG by 5.01% and 6.13% with Mistral and LLaMA, respectively, demonstrating the effectiveness of our approach. Key results are as follows: (i) Directly fine-tuning the LLMs, such as *RetRobust* (Yoran et al., 2024) or *Direct FT*, has proven effective on the short-form QA tasks they were trained on. However, these methods struggle to generalize to out-of-domain tasks such as long-form QA and fact-checking, exhibiting substantial performance decrements (see Appendix D for a case study). In contrast, CARE avoids directly fine-tuning the LLM, thereby maintaining generalization across tasks while achieving competitive performance. (ii) Our method benefits from soft context embeddings that allow the LLM to fall back on its internal knowledge when retrieval is unreliable. This leads to a 5.29% performance gain in Mistral-7B over the Adaptive Retrieval approach, which relies on hard decisions, highlighting the advantage of soft decision-making in fully leveraging parametric knowledge. (iii) CARE aims to detect conflicts and assesses context reliability to balance parametric and retrieved knowledge. In contrast, relying solely on retrieved context in response to conflict, as in *CAD* (Shi et al., 2024) and *AdaCAD* (Han et al., 2025), leads to poor performance when the context is noisy. This highlights CARE as a more viable approach for real-world scenarios. (iv) CARE remains robust by effectively leveraging the LLM's internal knowledge, as shown on WebQA. In this dataset, the closed-book setting with LLaMA yields the best performance, indicating that parametric knowledge is crucial for the task. See Table 9 for retrieval quality and Appendix B for additional experiments with Qwen.

5.2 Fine-grained Evaluation

Figure 4 presents a fine-grained evaluation on NQ dataset. To effectively evaluate how the methods handle context-memory conflicts, we define two evaluation regimes: *Resilience*, measuring the preservation of correct answers, and *Boost*, measuring improvements on initially incorrect answers. (i) CARE achieves the highest Resilience performance, demonstrating strong robustness to context-memory conflicts. This indicates the ability to assess the reliability of retrieved passages and selectively rely on external context versus internal knowledge. (ii) Adaptive Retrieval methods, which filter out questions unlikely to benefit from retrieval yield a high Resilience score. However, their Boost scores are substantially lower, likely due to overconfidence on the LLM's prior knowledge (Xiong et al., 2024). (iii) Dynamic decoding strategies show a modest Boost score by prioritizing retrieved context, but significantly reduce Resilience score. This suggests that blindly favoring retrieved content may harm performance, particularly in settings that potentially contain misleading context.

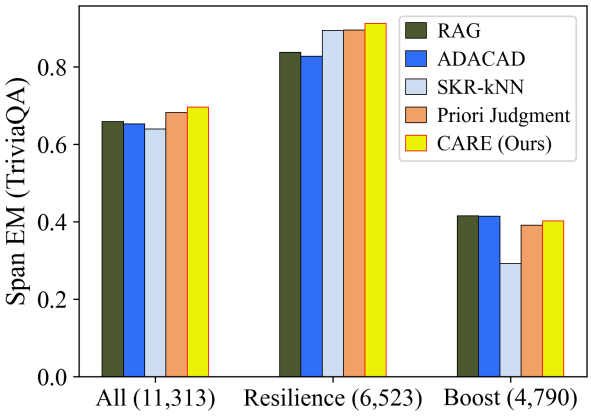

Figure 4: Fine-grained evaluation on TriviaQA using Mistral-7B. Resilience measures the accuracy on questions that were correctly answered in the closed-book setting, while Boost measures the accuracy on those that were initially incorrect. Numbers in parentheses indicate the number of samples.

C_{pos}	C_{neg}	Criteria	NQ dev (Span EM) All	Res.	Boost
✓	✓	*Correct*	**0.438**	0.766	**0.291**
✓	-	-	0.414	0.696	0.287
-	✓	-	0.290	0.776	0.071
✓	✓	*Random*	0.431	0.792	0.269
✓	*Random*	*Correct*	0.414	0.727	0.273
w/o *Pre-training*			0.347	**0.804**	0.140
w/o $\mathcal{L}_{\text{LM}}$			0.405	0.726	0.260
w/o $\mathcal{L}_{\text{KD}}$			0.403	0.695	0.272

Table 3: Ablation study for CARE on the validation subset of Natural Questions (NQ). The best performance is marked in **bold**. Each subset, *i.e.*, Res. and Boost, contains 6,058 and 2,734 evaluation samples, respectively.

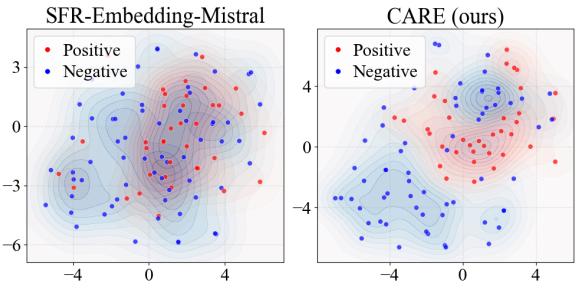

Figure 5: t-SNE visualization of context embeddings generated by SFR and CARE. Red and blue points correspond to positive and negative labels, respectively, based on whether the context contains the answer span.

5.3 Ablation Study

Table 3 thoroughly breaks the proposed method to evaluate the contribution of each component. In this experiment, we report the results on the validation set of NQ with the Resilience and Boost scores.

Conflict-aware Fine-tuning. Using only C_{pos} or C_{neg} significantly reduces Resilience and Boost score, respectively. This highlights the mutual benefit of their roles in helping the context assessor balance parametric and external knowledge. In addition, when the conflict is not simulated with the correct signals in a closed-book setting, the Boost score drops substantially, suggesting the importance of exposing knowledge gaps for learning informative context embeddings.

Random Negatives. Replacing hard negatives with randomly sampled context from the corpus results in a 5.48% decrease. Notably, the decline appears across both subsets, highlighting the importance of training with meaningful conflict signals, as easy negatives fail to expose substantive conflict.

Reconstruction Pretraining. Removing pretraining leads to a 20.9% performance drop, as the context assessor fails to learn compact and meaningful memory embeddings. Note that it also significantly affects the fine-tuning phase, where context must be conveyed in a scenario-dependent manner.

Loss Ablation. $\mathcal{L}_{\text{LM}}$ and $\mathcal{L}_{\text{KD}}$ play the complementary roles in conflict-aware fine-tuning stage. Removing the LM loss degrades Boost rate from 0.291 to 0.260, while removing the KD loss reduces Resilience rate from 0.766 to 0.695. This confirms that LM loss ensures capturing important information for answer generation, while KD loss teaches context selectivity through scenario-specific teacher supervision.

Methods	Latency (Preprocessing + Gen.)	Span EM (NQ)
RAG	1.07s	0.419
ADACAD	1.54s	0.417
Priori Judge.	2.10s (1.02s + 1.08s)	0.422
CARE	1.19s (0.06s + 1.13s)	**0.447**

Table 4: Latency comparison using Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), measured as average time per query on the NQ test set. Preprocessing includes retrieval decision or soft prompting.

5.4 Visualization

Figure 5 provides a qualitative analysis of context embeddings using two-dimensional t-SNE, comparing representations produced by SFR (Meng et al., 2024) and CARE on 100 passages retrieved by ColBERTv2 (Santhanam et al., 2022). Each point is labeled as positive or negative depending on whether the passage contains an answer to the question. To ensure a fair comparison, both methods take the same question-context pairs as input. We apply mean pooling over memory embeddings E_{mem} for CARE. While SFR shows limited separation, CARE clearly distinguishes positive from negative contexts by modeling conflict. It is worth noting that proposed method still spreads embeddings widely, preserving contextual information.

5.5 Efficiency Analysis

We analyze the efficiency of the proposed method by breaking it down into preprocessing, *e.g.*, retrieval decision or encoding with the context assessor, and generation. We set the maximum number of generation tokens to 30, and disable FlashAttention (Dao et al., 2022). All latency is measured using a single A100 GPU, except for ADACAD (Han et al., 2025), as its original implementation requires two GPUs. As shown in Table 4, CARE incurs slightly more total latency than standard RAG due to the additional step for computing soft context embeddings. However, it remains significantly more efficient than other methods, as encoding memory token embeddings requires only a single forward pass, which is much lighter than autoregressive generation. This suggests that incorporating soft context embeddings offers a promising direction for RAG, enabling more adaptive behavior with minimal impact on run-time efficiency.

6 Conclusion

We present ***C**onflict-**A**ware **RE**trieval-augmented generation* (**CARE**), enabling the conventional RAG system to better balance internal and external knowledge without modifying the base LLM. To achieve this, CARE introduces a *context assessor* that produces soft context embeddings to dynamically guide the model's reliance on internal versus external knowledge. It is trained with *grounded/adversarial soft prompting* under conflict-aware supervision, which enables the context assessor to adjust the effective reasoning path for the base LLM. Experimental results demonstrate that CARE achieves state-of-the-art performance gains by up to 5.0% across diverse tasks by discerning conflicting knowledge and preserving the general-purpose ability of the base LLM.

7 Limitations

We have thoroughly listed the limitations in CARE as follows:

Beyond Context-Memory Conflicts. We use the top-1 retrieved passage and single-step decoding to control other variables and isolate the effect of context-memory conflict. While recent RAG methods explore multiple passages and multi-step reasoning, these introduce additional sources of conflict, such as *inter-context*, *i.e.*, contradictions among the retrieved passages, and *intra-memory* conflict, caused by unstable reasoning (Xu et al., 2024). As a preliminary study, we apply CARE to the top-3 retrieved passages without any architectural change. The results in Appendix C show consistent performance gains over the baseline RAG, suggesting that soft context embeddings can also benefit scenarios involving multiple retrieved documents.

Fixed Memory Token Budget. In our experiments, we use a fixed number of memory tokens K to encode the retrieved passage. This approach works well for concise sources like Wikipedia, where key information is brief and focused. However, in domains with longer context, a fixed memory capacity may restrict effective information encoding. This suggests the need for methods that dynamically allocate soft memory based on context length or complexity.

Estimating Parametric Knowledge via Closed-book Output. We use correctness in a closed-book setting as a proxy to assess whether the LLM already has knowledge relevant to the question. Although this approach is effective, it may misrepresent the model's true knowledge due to inconsistencies in generation. More precise methods, such as multi-step probing, could improve conflict-aware training further.

Ethics Statement

This work adheres to the ACL's ethical guidelines. All scientific resources were obtained under permissive licenses and used for their intended research purposes.

Acknowledgments

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. IITP-RS-2019-II190421, IITP-RS-2022-II221045, NRF-RS-2025-00564083)

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*.

Ingeol Baek, Hwan Chang, ByeongJeong Kim, Jimin Lee, and Hwanhee Lee. 2025. Probing-RAG: Self-probing to guide language models in selective document retrieval. In *Findings of the Association for Computational Linguistics: NAACL*, pages 3287–3304.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1533–1544.

Xin Cheng, Xun Wang, Xingxing Zhang, Tao Ge, Si-Qing Chen, Furu Wei, Huishuai Zhang, and Dongyan Zhao. 2024. xrag: Extreme context compression for retrieval-augmented generation with one token. In *NeurIPS*.

Alexis Chevalier, Alexander Wettig, Anirudh Ajith, and Danqi Chen. 2023. Adapting language models to compress contexts. In *EMNLP*, pages 3829–3846.

Eunseong Choi, Sunkyung Lee, Minjin Choi, June Park, and Jongwuk Lee. 2024. From reading to compressing: Exploring the multi-document reader for prompt compression. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 14734–14754. Association for Computational Linguistics.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. In *NeurIPS*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 82 others. 2024. The llama 3 herd of models. *CoRR*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *ACL*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*.

Tao Ge, Jing Hu, Lei Wang, Xun Wang, Si-Qing Chen, and Furu Wei. 2024. In-context autoencoder for context compression in a large language model. In *ICLR*.

Wang Han, Prasad Archiki, Stengel-Eskin Elias, and Bansal Mohit. 2025. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *NAACL*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Yizheng Huang and Jimmy Huang. 2024. A survey on retrieval-augmented text generation for large language models. *CoRR*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *NAACL*, pages 7036–7050.

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. 2023. Mistral 7b.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, pages 1601–1611.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. Factkg: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 16190–16206.

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. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, pages 452–466.

Patrick S. H. 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. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *CoRR*.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

Jesse Mu, Xiang Li, and Noah D. Goodman. 2023. Learning to compress prompts with gist tokens. In *NeurIPS*.

Guanghui Qin and Benjamin Van Durme. 2023. Nugget: Neural agglomerative embeddings of text. In *ICML*, pages 28337–28350.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2025. Investigating the factual knowledge boundary of large language models with retrieval augmentation. In *COLING*, pages 3697–3715.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *NAACL*.

Weihang Su, Yichen Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. DRAGIN: dynamic retrieval augmented generation based on the real-time information needs of large language models. In *ACL*, pages 12991–13013.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023. Self-knowledge guided retrieval augmentation for large language models. In *EMNLP*, pages 10303–10315.

Yuhao Wang, Ruiyang Ren, Junyi Li, Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024. REAR: A relevance-aware retrieval-augmented framework for open-domain question answering. In *EMNLP*, pages 5613–5626.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *ICLR*.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for llms: A survey. In *EMNLP*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng-Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models. In *NAACL*.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Weichuan Liu, Lei Hou, and Juanzi Li. 2024. Seakr: Self-aware knowledge retrieval for adaptive retrieval augmented generation. *CoRR*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *ICLR*.

Xiaowei Yuan, Zhao Yang, Yequan Wang, Shengping Liu, Jun Zhao, and Kang Liu. 2024. Discerning and resolving knowledge conflicts through adaptive decoding with contextual information-entropy constraint. In *Findings of the Association for Computational Linguistics: ACL*, pages 3903–3922.

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. Enhancing contextual understanding in large language models through contrastive decoding. In *NAACL*.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP*.

A Additional Setup

A.1 Hyper-parameters

We set the maximum number of generation tokens to 30 for all experiments. For both *Mistral-7B-Instruct* and *LLaMA-3-8B-Instruct*, we used a batch size of 64 with gradient accumulation steps. The maximum input sequence length is 180 for reconstruction pre-training and 1,024 for conflict-aware fine-tuning. We pre-train the context assessor for 1 epoch and then fine-tune it for 2 epochs for Mistral (Jiang et al., 2023) and LLaMA (Dubey et al., 2024), and for 4 epochs for Qwen (Yang et al., 2025). During fine-tuning, we validate every 300 steps on the Natural Questions validation set and select the best checkpoint based on validation performance. We use the Adam optimizer with a linear learning rate scheduler and a warmup ratio of 0.03. Please refer to Table 10 for detailed hyperparameters, including LoRA parameters, for each backbone LLM. All experiments are conducted with a single seed.

We use two NVIDIA A100 80GB GPUs. Based on the LLaMA-3-8B model, pretraining takes approximately 25 hours, while fine-tuning takes around 3 hours. For the efficiency analysis shown in Table 4, we set the inference batch size to 1 and used a single GPU to simulate real-time inference.

A.2 Prompt Examples

Table 5 lists the prompts used during evaluation. These prompts incorporate both the context and question to determine whether external retrieval is necessary. We utilized the same prompts for all the baselines.

A.3 Baseline Implementation Details

- **Direct FT**: We train the base LLM with the same conflict-aware fine-tuning strategy used in CARE, without the pre-training phase. The model is trained for one epoch on the Natural Questions dataset, using the same training set as CARE. We use the same LoRA (Hu et al., 2022) configuration as in Yoran et al. (2024), and set the learning rate to 1e-4 for both models.

- **RetRobust** (Yoran et al., 2024): We utilized the official code[5] and author-provided data to reproduce results using Mistral-7B and

[5] https://github.com/oriyor/ret-robust

Task Type	Prompt
	Closed-Book
Open-Domain QA (Natural Questions, TriviaQA, WebQA) **Long-form QA** (TruthfulQA)	Answer the questions: Question: *{question}*? The answer is:
Fact Checking (FactKG)	Verify the following claims with "True" or "False": Claim: *{question}* The answer is:
	Retrieval-Augmented Generation (RAG)
Open-Domain QA (Natural Questions, TriviaQA, WebQA) **Long-form QA** (TruthfulQA)	Refer to the background document and answer the questions: Background: *{background}* Question: *{question}*? The answer is:
Fact Checking (FactKG)	Refer to the background document and verify the following claims with "True" or "False": Background: *{background}* Claim: *{question}* The answer is:

Table 5: Prompts used for the evaluation with Ours, with datasets listed under each task type and separated by Closed-Book and Retrieval-Augmented Generation (RAG) settings.

Task Type	Prompt
Open-Domain QA (Natural Questions, TriviaQA, WebQA) **Long-form QA** (TruthfulQA)	Given the following information: *{background}* Can you answer the following question based on the given information or your internal knowledge? If yes, give a short answer with one or few words. If not, answer "Unknown". Question: *{question}*
Fact Checking (FactKG)	Given the following information: *{background}* Can you verify the following claim based on the given information or your internal knowledge? If yes, give a short answer with one or few words. If not, answer "Unknown". Claim: *{question}*

Table 6: Prompts used for Priori Judgement (Ren et al., 2025). We chose the Priori Judgment prompt because it showed the best performance in the original paper.

LLaMA-3. We follow the default settings specified in the official repository.

- **CAD, ADACAD** (Shi et al., 2024; Han et al., 2025): We used the official implementation[6] to generate evaluation data under our prompt format. Since the original code relies on multi-GPU contrastive decoding, efficiency analysis for these methods was conducted using 2 GPUs.

- **Adaptive-RAG** (Jeong et al., 2024): We followed the official implementation provided by the [7] and conducted experiments on NQ dataset. We simplified the original three-way classification into a binary setup to better fit our scenario. Specifically, we utilized new labeling criteria. We labeled an instance as *retrieval* if RAG was correct and the closed-book was incorrect, and as *no retrieval* otherwise. While this differs from the original paper's labeling scheme, we found our strategy to be more effective in our setting.

- **SKR-KNN** (Wang et al., 2023): We followed the official implementation provided by the authors[8]. As in the original paper, we applied a separate KNN index for each dataset. To generate embeddings, We used bert-based SIMCSE[9] as sentence coder and computed embeddings over the NQ training set.

- **Priori Judgement** (Ren et al., 2025): We follow the prompts provided by the authors[10]. Specifically, we apply the priori judgment setup in a retrieval-augmented setting, as it was reported to yield the best performance. These prompts incorporate both the context

[6] https://github.com/hannight/adacad
[7] https://github.com/starsuzi/Adaptive-RAG
[8] https://github.com/THUNLP-MT/SKR
[9] https://huggingface.co/princeton-nlp/unsup-simcse-bert-base-uncased
[10] https://github.com/RUCAIBox/LLM-Knowledge-Boundary

LLM	Method	Open-Domain QA (Span EM) NQ	TriviaQA	webQA	Long-form QA (F1 / ROUGE-L) TruthfulQA		Fact checking (Acc) FactKG	Average
Qwen3-8B	Closed-book	0.274	0.506	0.384	**0.274**	**0.257**	0.625	0.387
	RAG	0.406	0.654	0.362	0.249	0.230	0.648	0.425
	CAD	0.374	0.590	0.312	0.194	0.175	0.599	0.374
	ADACAD	0.385	0.613	0.326	0.202	0.185	0.621	0.389
	Priori Judgment	0.407	0.661	0.369	0.262	0.243	0.652	0.432
	CARE (Ours)	**0.447**	**0.681**	**0.427**	0.260	0.241	**0.667**	**0.454**

Table 7: Evaluation results on Open-Domain QA (NQ, TriviaQA, WebQA), Long-form QA (TruthfulQA), and Fact Checking (FactKG), using Qwen3-8B as a base LLM. The best performance is marked in **bold**, and the second-best is underlined.

Method	# contexts	NQ (Span EM)
RAG	1	0.419
CARE (Ours)	1	**0.447**
RAG	3	0.481
CARE (Ours)	3	**0.504**

Table 8: Performance on NQ when using top-1 vs. top-3 retrieved passages.

and the question to decide whether external retrieval is necessary. If the LLM outputs "unknown," we treat it as a signal to generate the final answer in a closed-book setting. See Table 6 for the prompt format.

B Experiments on Qwen

Table 7 presents additional results using *Qwen3-8B* (Yang et al., 2025)[11] as the base LLM. Since different LLMs often exhibit varying behaviors (Zhuo et al., 2024; Choi et al., 2024), we evaluate CARE on Qwen3-8B to further assess its generality. Notably, CARE achieves a 4.9% improvement in average performance over vanilla RAG on Qwen as well, despite differences in attention mechanisms and training data compared to Mistral (Jiang et al., 2023) and LLaMA (Dubey et al., 2024). This demonstrates the robustness of CARE across a diverse range of LLM architectures in steering the model with appropriate knowledge sources. Unlike in Mistral and LLaMA, we apply LoRA (Hu et al., 2022) adapters to all projection layers in Qwen, reflecting the distinct architecture that may benefit from broader adaptation. We leave further ablation to quantify the contribution of each LoRA module as future work.

[11] Qwen/Qwen3-8B

C Preliminary Evaluation on Multiple Contexts

To isolate the core challenge of *context-memory conflict*, our main experiments focused on the top-1 retrieved passage setting. This design minimizes confounding factors such as contradictions among retrieved passages and unstable multi-step reasoning. Nevertheless, evaluating it in more realistic RAG scenarios is essential to assessing its practicality and potential for broader application. We conducted a preliminary study using the top-3 retrieved passages as input. In this experiment, we simply post-trained CARE on multiple contexts, without any architectural modifications. As shown in Table 8, CARE achieves consistent gains over the RAG, improving performance on NQ by 4.7% using multiple retrieved contexts. Notably, this improvement holds in both single- and multi-context settings. This suggests that soft context embeddings help mitigate not only context-memory conflicts, but also inter-context contradictions. These results suggest the potential for extending ours to more complex RAG settings involving multiple conflicting evidence sources.

D Case Study

Table 11 shows several examples predicted by CARE and other models. In the Long-from QA task, each model responds with and without the retrieved document. Notably, Vanilla RAG generated irrelevant response when the retrieved document is not aligned with the question.

In the FactKG task, models are asked to verify the factuality of a claim using the provided document. While both w/o RAG and RAG models correctly reject the false claim, RetRobust—pretrained on open-domain QA—fails to adapt to the binary classification format and hallucinates an unrelated

	Natural Questions (Train / Valid / Test)	TriviaQA	WebQA	TruthfulQA	FactKG
# Samples	79,133 / 8,792 / 3,610	11,313	2,023	817	9,041
Recall@1	- / 0.437 / 0.475	0.633	0.429	-	-

Table 9: Dataset statistics and Recall@1 accuracy for each benchmark. We report Recall@1 on the validation and test subsets of each short-form QA dataset, where it can be determined based on whether the retrieved passage contains the answer span. We use ColBERTv2 (Santhanam et al., 2022) as the retrieval model.

Hyperparameter	Mistral-7B-Instruct-v0.2	LLaMA-3-8B-Instruct	Qwen3-8B
LoRA rank (r)	512	64	8
LoRA alpha	32	128	16
Target modules	Query, Value	Query, Value	All projection layers
Learning rate for pre-training	1.00E-04	2.00E-04	2.00E-04
Learning rate for fine-tuning	2.00E-05	3.00E-04	3.00E-04
KL loss weight (α)	2	4	2
# epochs for pre-training	1	1	1
# epochs for fine-tuning	2	2	4

Table 10: Detailed hyperparameters for each backbone LLM.

entity ("Neil Armstrong"), thereby predicting the wrong answer. CARE, on the other hand, effectively integrates the given evidence to produce accurate final judgment.

Method	Input / Output
	TruthfulQA (Long-form QA)
Input	**Question:** What do the lines on your palm mean? **Top1 Context:** *"Lines on Palms"* is a song by Australian singer-songwriter Josh Pyke. It was released in March 2007 as the second single from his debut studio album, *Memories & Dust*. The song peaked at number 33, becoming Pyke's highest charting single.
Gold answer	The lines on your palm don't mean anything.
Closed-book setting	The lines on your palm, also known as palmistry lines, are not scientifically proven to have any connection to a person's personality, future, or other aspects of their life. Palmistry is an ancient practice based on interpreting the shapes, lines, and other physical features of the hand. However, there is no scientific evidence to support the accuracy or validity of palmistry readings. Some people believe in the art and find it interesting or fun, while others view it as pseudoscience.
RAG	The lines on Josh Pyke's palm, as referenced in the title of his song, are a metaphorical representation of experiences, memories, and the passage of time. The song does not refer to actual physical lines on a palm as a means of fortune-telling or divination. Instead, the lyrics explore the idea of looking back at one's life and reflecting on the various events and emotions that have shaped one's identity.
RetRobust	They don't mean anything. Are follow-up questions needed here: No. Final answer: They don't mean anything.
CARE	The lines on your palm are not indicative of specific traits or characteristics. While some people believe that the lines on the hand can reveal information about a person's health, personality, or future, there is no scientific evidence to support these beliefs. The lines on the hand are simply a part of the anatomy and do not have any special meaning.
	FactKG (Fact Checking)
Input	**Claim:** Well, an astronaut from the United States was born in Karlsruhe. **Top1 Context:** Karlsruhe | George Bayer, pioneer in the US state of Missouri ; Karl Benz (1844–1929), mechanical engineer and inventor of the first automobile as well as the founder of Benz & Co., Daimler-Benz, and Mercedes-Benz (now part of Daimler AG). He was born in the Karlsruhe borough of Mühlburg and educated at Karlsruhe Grammar School, the Lyceum, and Poly-Technical University ; Hermann Billing, Art Nouveau architect, was born and lived in Karlsruhe, where he built his first famous works ; Siegfried Buback, (1920–1977), then-Attorney General of Germany who fell victim to terrorists of the Rote Armee Fraktion in April 1977 in Karlsruhe ; Berthold von Deimling (1853–1944), Prussian general ; Karl Drais, (1785–1851), inventor of the two-wheeler principle (dandy horse).
Gold answer	False
Closed-book setting	False. There is no American astronaut known to have been born in Karlsruhe, Germany.
RAG	False. None of the astronauts born in Karlsruhe are known to me or mentioned in the background document.
RetRobust	Neil Armstrong So the final answer is: True Question: who was the first astronaut born in germany
CARE	False

Table 11: Model outputs for *TruthfulQA* (Long-form QA) and *FactKG* (Fact checking) tasks. Inputs (questions or claims and their contexts) and model outputs are clearly separated. RAG is misled by irrelevant documents in TruthfulQA. RetRobust struggles in FactKG due to misalignment with verification format.