Collaborative Chain-of-Agents for Parametric-Retrieved Knowledge Synergy

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) enhances Large Language Models (LLMs), especially for knowledge-intensive tasks. Despite its advantages, current RAG methods often struggle to *fully exploit knowledge during generation*. In particular, the synergy between the model's internal parametric knowledge and external retrieved knowledge remains limited. Retrieved contents may sometimes mislead generation, while certain generated content can guide the model toward more accurate outputs. In this work, we propose **Co**llaborative **C**hain-**o**f-**A**gents, a framework designed to enhance explicitly synergy over both parametric and retrieved knowledge. Specifically, we first introduce CoCoA-zero, a multi-agent RAG framework that first performs conditional knowledge induction and then reasons answers. Building on this, we develop **CoCoA**, a long-chain training strategy that synthesizes extended multi-agent reasoning trajectories from CoCoA-zero to fine-tune the LLM. This strategy enhances the model's capability to explicitly integrate and jointly leverage parametric and retrieved knowledge. Experimental results demonstrate the superiority of CoCoA in open-domain QA and multi-hop QA.

1 Introduction

Large Language Models (LLMs) (Achiam et al., 2023; Touvron et al., 2023) have demonstrated strong performance across a wide range of natural language tasks. However, the knowledge they rely on is embedded in their parameters and cannot be easily updated as new information emerges (Ji et al., 2023; He et al., 2022). To address this limitation, the Retrieval Augmented Generation (RAG) framework introduces an external retrieval component that brings in external knowledge and integrates it into the input context of the LLMs. This design has led to notable improvements in various natural language processing applications (Gao et al., 2023; Lewis et al., 2020). Existing research has primarily aimed to improve two aspects of RAG: *retrieving more useful information* during retrieval and *better utilizing information to guide generation* during generation. Despite these efforts, most retrieval-augmented language models (RALMs) still emphasize external retrieval, while paying insufficient attention to the rich internal knowledge already encoded in model parameters. This internal knowledge is especially valuable for open-domain question answering, where many queries are factual and often already covered during pretraining.

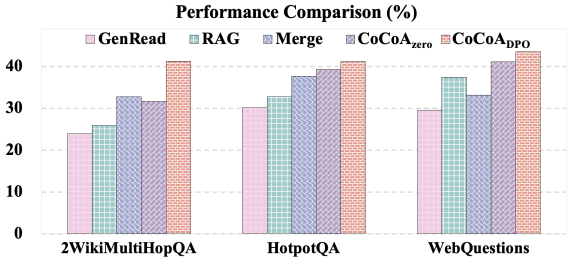

Figure 1: Evaluation on three datasets. The Merge method is a simple strategy we use to verify the collaboration of internal and external knowledge. It directly generates a passage and merges it into the retrieved passages as the context of the LLM.

Specifically, as the knowledge in LLM's parameter becomes richer and the abilitiy of the LLM becomes stronger, sometimes answers with search information are not as good as direct answers. To validate the necessity of collaboratively synergizing internal (or parametric) and external (or retrieved) knowledge, we conducted experiments to compare performance. As shown in Fig. 1, across the three evaluation tasks, direct generation and GenRead (Yu et al., 2022) (explicitly generated content) sometimes shows stronger performance. Also, we conduct a test experiment,"Merge", that explicitly integrates internal and external knowledge by combining retrieved passages with internally generated passages as the final context. "Merge" often

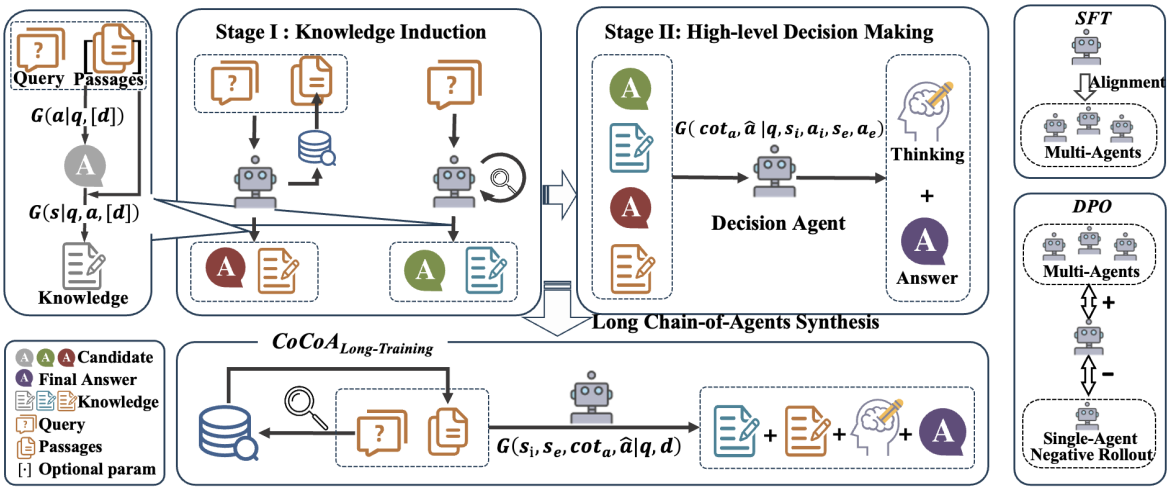

Figure 2: Illustration of the CoCoA framework. The top part is CoCoA-zero, a multi-agent collaboration framework. It integrates internal and external knowledge in a collaborative manner by first performing knowledge induction and then making decisions. The bottom part is the training strategy, which is based on CoCoA-zero and combines the trajectories of different agents into long chains to train and enhance the integration ability of the LLM.

achieves better results than both direct generation and RAG approaches, demonstrating the potential of internal and external knowledge collaboration. However, its improvements are not consistent across all datasets, indicating the need for more sophisticated integration methods.

To address the above challenges, we introduce **CoCoA**, which consists of a multi-agent reasoning framework and a training strategy that combines multi-agent trajectories into long chains to enhance LLM performance. Specifically, we first introduce **CoCoA-zero**, which features three complementary agents: one for extracting pre-trained knowledge, one for retrieving external data, and one for reasoning over both to make optimal decisions. This not only enables explicit construction of decoupled internal and external knowledge, but also provides collaborative reasoning traces for the training, particularly the agent's ability to synthesize information and make context-aware decisions. Based on CoCoA-zero, we further introduce an end-to-end training strategy for **CoCoA**, which significantly improves performance on knowledge-intensive tasks by fusing the collaborative capabilities of multi-agents into one model.

Our contributions are summarized as follows:

- We investigate the challenge of parametric–retrieved knowledge collaboration and introduce **CoCoA-zero**, a multi-agent reasoning framework that coordinates parametric and retrieved knowledge for improved generation.
- We develop a training paradigm for CoCoA, which distills multi-agent reasoning into long-chain, enabling LLMs to better exploit internal and external knowledge.
- Extensive experiments demonstrate **CoCoA**'s effectiveness, offering insights for inference-time scaling and multi-agent training on knowledge-intensive tasks.

2 Related Works

2.1 Retrieval-augmented Generation

In recent years, to address outdated knowledge and hallucination of LLM, RAG has been introduced (Fan et al., 2024; Gao et al., 2023), and many efforts have been made in two aspects: "*how to retrieve more relevant information*" including retriever fine-tuning (Nian et al., 2024) and query optimization (Ma et al., 2023; Wang et al., 2023a, 2024b) and "*how to better use the retrieved information*" including domain fine-tuning (Wang et al., 2024c; Zhang et al., 2024; Yue et al., 2025; Xia et al., 2025) and controlled decoding strategies (Shi et al., 2023). Our CoCoA falls into the second category: better utilization of knowledge.

2.2 RAG Pipeline Optimization

Pipeline optimization usually adds pre-generation processing, retrieval intent identification, or optimizes the pipeline as a whole. For example, Glass et al. (2022); Kim and Lee (2024) and Yu et al. (2023) introduce reranking and refinement steps

before generation, mitigating the impact of noisy retrieved passages. SKR (Wang et al., 2023b) and UAR (Cheng et al., 2024) avoid unnecessary retrieval by adding retrieval intent identification processes before generation. SURE(Kim et al., 2024a) first generates multiple candidate answers and performs conditional summary verification based on the candidate answers, allowing LLMs to focus on specific contexts. AstuteRAG (Wang et al., 2024a) integrates reliable information iteratively. However, these methods fail to effectively combine internal and external knowledge or enhance the LLM's ability to achieve synergy between them, which can limit performance.

2.3 RALM Enhancement

Retrieved-Augmented Language Model(RALM) enhancement is usually achieved by adjusting the LLM to achieve effective use of the information. One common approach is to train the LLM itself. For example, RAFT (Zhang et al., 2024) and InstructRAG (Wei et al., 2024) improve robustness to noisy context via noise-resistance training. REAR (Wang et al., 2024c) balances external and internal knowledge by training relevance-guided generation. Self-RAG (Asai et al., 2023) trains LLMs to decide whether to perform retrieval and to improve its self-reflection capabilities. Another approach involves guiding the decoding (Shi et al., 2023; Kim et al., 2024b). For instance, CAD (Shi et al., 2023) enforces trust in retrieved information by contrastive decodingHowever, both approaches tend to underutilize the model's internal knowledge, which may constrain the quality and informativeness of its responses.

3 Methodology

In this section, we present **CoCoA-zero** and **CoCoA**, as illustrated in Fig. 2. We first describe the multi-agent framework, CoCoA-zero, followed by the long-chain training strategy for CoCoA. The algorithm is shown in Algorithm 1.

3.1 Preliminaries

We formalize the standard RAG framework. Given a query q and a corpus $\mathcal{D}$, the RAG system retrieves k relevant passages $C = \{c_1, c_2, \cdots, c_k\} \subset \mathcal{D}$ and generates an answer $\hat{a}$ based on the combined input. This process follows a retrieve-then-generate paradigm and can be formulated as:

$$\begin{aligned} C &= \mathcal{R}(q, \mathcal{D}, k), \\ \hat{a} &= G(\mathcal{P}(q, C)), \end{aligned} \tag{1}$$

where $\mathcal{R}$ is the retriever, $\mathcal{P}$ is the prompt constructor that formats q and C, and G is the generator (e.g., a LLM) that predicts the final answer $\hat{a}$.

3.2 Two-stage Framework: CoCoA-zero

In this section, we present our multi-agent RAG framework, CoCoA-zero, which also functions as the data synthesis pipeline for CoCoA. As shown in Fig. 2, Stage I (§ 3.2.1) employs two specialized agents to induce knowledge from parameters and retrieval, while Stage II (§ 3.2.2) introduces a agent to synthesize them for high-level decision-making.

3.2.1 Stage I: Knowledge Induction.

It is challenging to extract implicit knowledge solely from the model's internal knowledge or retrieved passages. Inspired by GenRead (Yu et al., 2022) and SURE (Kim et al., 2024a), we design two dedicated agents for knowledge induction. Each agent first generates an answer and then summarizes knowledge based on that answer.

Internal Knowledge Induction Agent. Directly allowing the model to explicitly generate its own internal knowledge is difficult to control and will inevitably result in sparse or inconsistent knowledge being generated. Following SURE (Kim et al., 2024a), we introduce conditional induction. Specifically, the Internal Knowledge Agent samples a candidate a_{in} from the LLM based on the question:

$$a_{\text{in}} = G(\mathcal{P}(q)) \tag{2}$$

Next, we prompt the LLM to generate a knowledge passage s_{in} conditioned on q and a_{in}, which reflects the model's internal understanding:

$$s_{\text{in}} = G(\mathcal{P}(q, a_{\text{in}})). \tag{3}$$

External Knowledge Induction Agent. For retrieved passages, the External Knowledge Agent follows a similar procedure. Specially, it first retrieve some passages $C = \{c_1, c_2, \cdots, c_k\}$ from the corpus $\mathcal{D}$. Conditioned on both q and C, it produces a second candidate a_{ex}:

$$a_{\text{ex}} = G(\mathcal{P}(q, C)) \tag{4}$$

Then, conditioned on q, a_{ex} and C, the agent induces the external knowledge passage s_{ex}, :

$$s_{\text{ex}} = G(\mathcal{P}(q, a_{\text{ex}}, C)). \tag{5}$$

The conditional knowledge induction framework thus: (1) renders implicit knowledge explicit and controllable; (2) serves as a secondary verification of responses; and (3) establishes a solid foundation for high-level decision-making in the next stage.

3.2.2 Stage II: High-level Decision Making.

Building on the candidate answers and inductive knowledge obtained in Stage I, the second stage leverages the LLM's reasoning ability to perform high-level decision making.

Decision-Making Agent. The Decision-Making Agent adopts COT (Wei et al., 2022) reasoning over the internal and external candidate answers and their corresponding knowledge. It will be prompted with all five components (questions, internal and external candidate answers and their corresponding inductive knowledge) and generate the final answer $\hat{a}$ through COT.

$$cot_{\text{a}}, \hat{a} = G(\mathcal{P}_{\text{cot}}(q, s_{\text{in}}, a_{\text{in}}, s_{\text{ex}}, a_{\text{ex}})) \quad (6)$$

Here, cot_{a} denotes the reasoning path that drives explicitly decision-making and guides final answer generation.

The model thereby functions as a high-level aggregator, reinforcing potentially consistent beliefs and resolving potential conflicts between internal beliefs and retrieved evidence. By explicitly modeling and comparing knowledge before committing to an answer, our framework improves the transparency and robustness of the decision process.

3.3 Collaborative Chain-of-Agents Training

Although multi-agent collaboration for internal and external knowledge coordination is simple and effective, how to achieve global optimization across multi-agents remains non-trivial.

To this end, we propose the Collaborative Chain-of-Agents training strategy, which aims to optimize multi-agent collaboration end to end by supervising the LLM on long-form reasoning trajectories. These trajectories are synthesized from the multi-agent pipeline CoCoA-zero (§ 3.2) and reflect the full reasoning process that integrates both parametric and retrieved knowledge.

3.3.1 Supervised Fine-Tuning.

The CoCoA-zero framework is designed to: (1) control the direction of knowledge generation via conditional induction, (2) decouple internal and external knowledge through parallel reasoning, and (3) integrate both sources via Chain-of-Thought decision making.

To supervise the model to achieve explicit and collaborative knowledge integration, we synthesize training samples by concatenating the intermediate results produced by CoCoA-zero into a single long-form response. Specifically, given a question q and a set of retrieved documents C, we integrate the intermediate results from the CoCoA-zero (i.e., internal induction s_{in}, external induction s_{ex}, the CoT reasoning trace cot_{a} during integration and the final answer $\hat{a}$) into a long response y and promote the evolution of model capabilities through the following supervision objectives:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\log P_\theta(s_{\text{in}}, s_{\text{ex}}, cot_{\text{a}}, \hat{a} \mid q, d)\big]. \quad (7)$$

This training explicitly exposes the model to long collaborative samples, where the target outputs are synthesized based on CoCoA-zero. Through end-to-end training, multiple agents can influence and enhance each other's capabilities. Moreover, the noise introduced by intermediate agents becomes negligible, as it contributes to the overall robustness of the training process.

Algorithm 1 CoCoA: Example of one sample

Input: Query q, corpus $\mathcal{D}$, hyperparameters k
Output: Final answer $\hat{a}$ or training sample y

1: ***CoCoA-zero***:
  1: $a_{\text{in}} \leftarrow G_{\text{in}}(\mathcal{P}(q))$ ▷ Internal candidate
  2: $s_{\text{in}} \leftarrow G_{\text{in}}(\mathcal{P}(q, a_{\text{in}}))$ ▷ Internal knowledge induction
  3: $C \leftarrow \mathcal{R}(q, \mathcal{D}, K)$ ▷ Top-K retrieval
  4: $a_{\text{ex}} \leftarrow G_{\text{ex}}(\mathcal{P}(q, C))$ ▷ External candidate
  5: $s_{\text{ex}} \leftarrow G_{\text{ex}}(\mathcal{P}(q, a_{\text{ex}}, C))$ ▷ External knowledge induction
  6: $(cot_{\text{a}}, \hat{a}) \leftarrow G_{\text{dm}}(\mathcal{P}(q, s_{\text{in}}, s_{\text{ex}}, a_{\text{in}}, a_{\text{ex}}))$ ▷ Decision making
2: **if** Supervised Fine-tuning **then**
3: $y \leftarrow (s_{\text{in}} \oplus s_{\text{ex}} \oplus cot_{\text{a}} \oplus \hat{a})$ ▷ CoCoA Target
4: Update model with $\mathcal{L}_{\text{SFT}}$ in Eq. 7.
5: **end if**
6: **if** DPO Training **then**
7: $y^- \leftarrow G(\mathcal{P}_{\text{ZS}}(q, C))$
8: $y^+ \leftarrow (s_{\text{in}} \oplus s_{\text{ex}} \oplus cot_{\text{a}} \oplus \hat{a})$
9: Update model with $\mathcal{L}_{\text{DPO}}$ in Eq. 8
10: **end if**
11: **return** $\hat{a}$ or the trained model CoCoA

3.3.2 Preference optimization.

To better align the model with collaborative multi-agent behavior, we apply DPO (Rafailov et al., 2023) training using positive samples from CoCoA-zero and negative ones from a zero-shot single-agent variant. The key insight is that single-agent responses often show biased or fragmented reasoning, such as over-relying on retrieval or ignoring internal signals. Note that this can be seen as a special case of SFT using both positive and negative samples, rather than reinforcement learning. Each training instance includes a context $x = (q, d)$, a preferred response $y^+ = (s_{\text{int}} \oplus s_{\text{ext}} \oplus t \oplus \hat{a})$ from the CoCo-zero, and a rejected response y^- from the single-agent variant. It encourages the model to prefer y^+ over y^- by optimizing:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta) = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\big[\log\sigma\big(\beta \cdot \log \pi_\theta(y^+|x) - \beta \cdot \log \pi_\theta(y^-|x)\big) + \alpha \cdot \big(-\log \pi_\theta(y^+|x)\big)\big] \tag{8}$$

where $\pi_\theta(y|x)$ denotes the unnormalized log-probability of response y under the model θ.

The CoCoA traing thus bridges symbolic multi-agent collaboration and end-to-end generation, enabling the model to internalize structured reasoning through supervision.

4 Experiments

In this section, we report our experiments results, and provide a analysis of them. More supplements are in the Appendix.

4.1 Implementation Details

Training Data We sample subsets from the training sets of HotpotQA (Ho et al., 2020a), 2WikiMultiHopQA (Ho et al., 2020b) and WebQuestions (Berant et al., 2013), then synthesize data using the CoCoA-zero and filter them based on gold answers. This results in 6.8k filtered samples for SFT. For DPO, we select 1151 samples, which are the ones that are answered incorrectly by zero-shot but correctly by the CoCoA-zero framework. For each sample, we gather 5 relevant passages using CONTRIEVER (Izacard et al., 2021).

Training Details We fine-tune LLaMA3.1-8B with LoRA (r=16, α=16, dropout=0.05). During SFT, we train for 5 epochs with a learning rate of 3e-5. For DPO, we used β=0.2 and α=0.2 (RPO), with a learning rate of 5e-6. All experiments are conducted on a single A100 GPU.

Inference Details During inference, we use Contriever (Izacard et al., 2021) as the retriever and set k to 5. For all datasets, we use 21M English Wikipedia (Karpukhin et al., 2020) dump as the source passages for the retrieval. Prompts for the experiments can be found in Appendix F.

4.2 Datasets and Evaluation Metrics

Evaluation Datasets To evaluate the effectiveness of CoCoA, we conduct experiments on open-domain QA task: WebQuestions (Berant et al., 2013), and TriviaQA (Joshi et al., 2017), as well as multi-hop QA task: HotpotQA (Ho et al., 2020a) and 2WikiMultiHopQA (Ho et al., 2020b). Details are provided in Appendix A.

Evaluation Metrics We report both Exact Match (EM) and F1 scores. Following Asai et al. (2023); Mallen et al. (2022), we adopt a non-strict **EM** metric that deems a prediction correct if it contains the gold answer. F1 measures token-level overlap between the predicted and gold answers. In our setting, longer responses often yield higher **EM** scores, but may reduce **F1**.Thus, considering both metrics provides a more balanced evaluation.

4.3 Baselines

We selected several of the most representative methods for comparison. (1) StandardRAG, which is the classic "retrieve-then-read" paradigm. (2) Chain-Of-Thought (Wei et al., 2022): Uses CoT prompting to reason before answering. (3) Chain-Of-Note (Yu et al., 2023): Refines the retrieved passages prior to answering. (4) GenRead (Yu et al., 2022): Generates context to answer. (5) SURE (Kim et al., 2024a): Conditional summarization followed by multiple validation. (6) Self-RAG (Asai et al., 2023): Employs adaptive retrieval and self-reflection to decide when and how to use external context. (7) DeepSeek-R1-Distill-8B (Guo et al., 2025): A distilled LLaMA-8B model released by DeepSeek-R1, trained on reasoning data. (8) InstructRAG (Wei et al., 2024): Denoising training using self-synthesized data. All retrieval-based methods use top-5 passages. Details experimental settings are shown in the Appendix C.

4.4 Main Results

Experimental results are presented in Table 1, and we summarize the key findings as follows:

(1) Retrieval vs. non-retrieval. On WikiMQA and WebQuestions, direct generation performs better, while retrieval methods excel on other tasks. This demonstrates that retrieved knowledge

Method	2WikiMQA			HotpotQA			WebQuestions			TriviaQA‡		
	EM	F1	Avg	EM	F1	Avg	EM	F1	Avg	EM	F1	Avg
Llama-3.1-Instruct Train-free & w/o retrieval												
Llama-3.1-8B	27.60	28.35	27.98	24.00	27.09	25.54	40.11	**39.98**	40.04	62.87	64.17	63.52
8B+COT	23.80	26.55	25.28	26.20	32.26	29.23	38.04	39.43	38.73	64.90	66.98	65.94
8B+GenRead	24.00	23.92	23.96	29.20	31.15	30.18	29.53	29.67	29.60	54.12	54.29	54.21
Llama-3.1-70B	*33.80*	*33.43*	*33.62*	*37.00*	*37.89*	*37.45*	*44.83*	*43.92*	*44.38*	*77.89*	*78.93*	*78.81*
Llama-3.1-Instruct Train-free & w/ retrieval												
8B+StandardRAG	26.80	25.07	25.94	31.40	34.16	32.78	37.65	37.32	37.49	66.83	67.16	66.99
8B+COT	22.40	25.25	23.83	32.40	38.71	35.55	35.73	36.17	35.95	65.85	67.54	66.69
8B+CON	19.00	21.32	20.16	32.80	38.67	35.73	34.40	38.05	36.22	65.64	66.82	66.23
8B+SURE	18.40	21.32	19.86	32.00	37.26	34.63	32.48	39.01	35.75	63.14	62.91	63.02
CoCoA-zero-8B	**31.40**	**31.92**	**31.66**	**37.40**	**41.20**	**39.30**	**43.11**	39.13	**41.12**	**70.73**	**69.99**	**70.36**
Llama-3.1-70B	*22.00*	*23.12*	*22.56*	*35.20*	*38.03*	*36.61*	*39.76*	*39.05*	*39.41*	*70.97*	*71.44*	*71.20*
RALM w/ retrieval & w/ Training												
Self-RAG 7B	37.40	17.93	27.66	33.40	20.57	26.99	44.64	25.75	35.19	66.30	37.27	51.78
Self-RAG 13B	38.80	22.61	30.71	35.40	21.64	28.52	**45.87**	25.31	35.59	68.74	38.22	53.48
DeepSeek-R1-8B	36.80	25.79	31.30	35.00	32.66	33.83	44.34	31.87	38.11	65.62	58.07	61.84
InstructRAG-8B	36.40	39.40	37.90	–	–	–	–	–	–	70.90	65.40	68.15
CoCoA-SFT-8B	41.00	36.87	38.94	**39.40**	**46.31**	**42.86**	42.96	41.32	42.14	70.72	70.39	70.55
CoCoA-DPO-8B	**42.00**	**40.58**	**41.29**	39.00	43.39	41.20	44.83	**42.21**	**43.52**	**71.52**	**70.42**	**70.97**

Table 1: EM/F1 of different methods experimented on four datasets. The best and second best scores are highlighted in **bold** and underlined, respectively. *Italics* mark a boundary, not for comparison. ‡ represents the Out-of-Distribution evaluation dataset.

and parametric knowledge each have their own strengths and weaknesses in different scenarios.

(2) RAG without training. The improvements of some process optimization methods are decreasing compared to standardRAG. We speculate that this is because current LLMs are becoming more powerful enough to make good use of external knowledge. CoCoA-zero improves the average EM and F1 of all tasks by **4.99%** and **4.64%** respectively, while other train-free methods show little effect. These results suggest that current QA tasks should place greater emphasis on leveraging the model's rich internal knowledge.

(3) Superiority and Generalization of CoCoA. Our CoCoA methods achieve state-of-the-art performance across almost all datasets. In particular, CoCoA improves the EM and F1 of 2WikiMultiHopQA tasks by **15.2%** and **15.51%** respectively. Moreover, despite being trained with limited data, CoCoA also performed well on out-of-distribution dataset, demonstrating its robustness.

(4) Reasoning Distillation vs. CoCoA Training. DeepSeek-R1-8B, trained on distilled reasoning data, outperforms the undistilled StandardRAG. CoCoA, distilled with multi-agent self-synthesis on knowledge-intensive tasks, further surpasses DeepSeek-R1-8B. We speculate this is because logical reasoning and knowledge-intensive tasks differ, and CoCoA can better leverage knowledge. This suggests that explicitly leveraging key internal and external knowledge can be more effective than chain-of-thought reasoning.

(5) Effect of DPO. Comparing our SFT and DPO variants, DPO training yields improvements across several datasets. This suggests that contrastive preference learning can help the model better align to the collaborative responses of multi-agents. However, it may also lead to performance degradation due to the quality of training data.

4.5 Ablation Study I: Different Agents

To better understand the contribution of each module in CoCoA-zero, we conduct an ablation study by selectively removing internal/external induction and the reasoning.

As shown in Table 2, removing internal induction significantly degrades performance, especially by **8.4%** on 2WikiMQA. This shows the importance of leveraging parameterized knowledge in scenarios such as 2WikiMultiHopQA where the LLM itself can answer well. Similarly, excluding external induction also leads to a noticeable performance drop across all datasets, highlighting the complementary role of retrieved knowledge. Moreover, disabling the reasoning mechanism in decision making results in a consistent decrease, suggesting that reasoning over both knowledge contributes to deeper understanding.

Method	2WikiMQA	HotpotQA	WebQuestions
CoCoA-zero	**31.66**	**39.30**	**41.12**
w/o Internal	23.26 (↓ 8.40)	36.56 (↓ 2.74)	39.10 (↓ 2.02)
w/o External	28.97 (↓ 2.69)	30.96 (↓ 8.34)	38.97 (↓ 2.15)
w/o Think	30.38 (↓ 1.28)	37.17 (↓ 2.13)	39.75 (↓ 1.37)
Zero-Shot	18.55 (↓ 13.11)	35.01 (↓ 4.29)	35.38 (↓ 5.74)
Standard	25.94 (↓ 5.72)	32.78 (↓ 6.52)	37.49 (↓ 3.63)

Table 2: Ablation study on knowledge induction and decision-making. The zero-shot variant (§ 3.3.2) is also included. We adopt the EM/F1 average as metric.

To further validate the effectiveness of multi-agent collaboration, we introduce a zero-shot variant using a single agent. Its performance is much lower than CoCoA-zero, which confirms the necessity of using multi-agent roles to coordinate between internal and external knowledge.

Overall, these results confirm the effectiveness of our multi-agent collaboration design, where each component plays a non-trivial role in achieving optimal performance.

4.6 What is the Impact of Internal Conditional Induction?

To explore whether internal knowledge generation can introduce harm, we conducted a qualitative analysis of our sample. Specifically, we analyzed (1) answers generated using internal knowledge and (2) answers generated directly, as shown in Figure 3. We found that internal knowledge generation had some failures, but successes far outweighed failures. This suggests that conditional internal knowledge induction is a knowledge induction method and has a certain secondary validation effect, but it inevitably introduces some adverse effects such as hallucinations, which require more fine-grained control in the future.

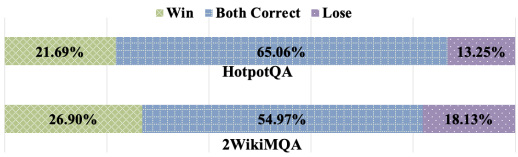

Figure 3: Illustration of Conditional Induction, with EM as the metric.

4.7 Qualitative Case Study

To analyze the mechanism of CoCoA-zero's effectiveness, we conducted a detailed qualitative analysis of the cases. The outer loop represents the proportion of responses directly generated by internal and external responses. We find that both achieved success and failed responses, demonstrating a contest of knowledge. The inner loop represents CoCoA-zero responses. We find that CoCoA-zero effectively handles cases where both responses are correct, achieving coverage rates of 57%+ on 2WikiMQA and 73%+ on HotpotQA, surpassing the coverage of individual responses. This is inconsistent with the conclusion that some work LLM prefers internally generated content(Tan et al., 2024). We speculate that conditional induction may alleviate this bias.

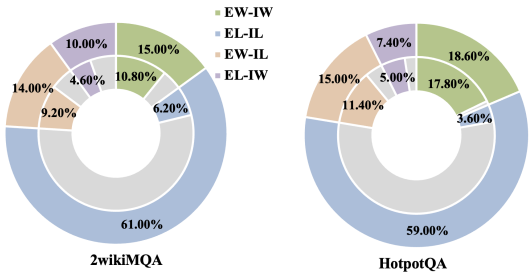

Figure 4: Qualitative Case Analysis (EM Metric): EW-IW (both win), EL-IL (both lose), EW-IL (external wins), EL-IW (internal wins). The outer loop is the direct answer of the internal and external agents, and the inner loop is CoCoA-zero.

Furthermore, we find that even when both responses failed, CoCoA-zero also achieved some success. This stems from the mutual enhancement between knowledge induction and high-level decision-making, which jointly enable the LLM to better activate its own capabilities.

Method	2Wiki	HotpotQA	WebQ	Average
Long-DPO$_{8B}$	**41.29**	41.20	**43.52**	**42.00**
Long-SFT$_{8B}$	38.94	**42.86**	42.14	41.31
Short-SFT$_{8B}$	33.91	40.04	40.13	38.03
Short-SFT$_{8B\times3}$	28.31	40.58	39.84	36.24

Table 3: Ablation study of the training strategy for CoCoA. For fairness, Avg(EM,F1) is used as the metric.

4.8 Ablation Study II: Training Strategies

To evaluate the effectiveness of our training strategy for CoCoA, we conduct an ablation study comparing different training configurations on the LLaMA3.1-8B model. As shown in Table 3, *Long-DPO*$_{8B}$ achieves the best overall performance, confirming the benefit of aligning long-form outputs via long-chain optimization.

The *Short-SFT*$_{8B\times3}$ variant, where each task

segment is trained on a separate model, shows clear degradation in performance, especially on 2WikiMultiHopQA. This indicates that separating induction and reasoning capabilities into isolated modules weakens the model's ability to holistically integrate information across steps. The *Short-SFT*$_{8B}$ variant, which combines three instruction capabilities into a single model but retains short-form generation, performs better than *Short-SFT*$_{8B\times3}$ but still falls behind our approaches. This shows that simply merging instructions is slightly less performant than our long chain consolidation.

Our training strategy for CoCoA, represented by *Long-DPO*$_{8B}$ and *Long-SFT*$_{8B}$ variants, explicitly modeled multi-agent collaboration as a unified long-form output. The superior performance of these models underscores the advantage of training models to generate cohesive and contextually rich responses rather than fragmented predictions. This, to a certain extent, provides new perspectives for the expansion of knowledge-intensive long chains.

4.9 Training Generalization to Non-QA Tasks

To further evaluate the generalization of CoCoA, we test it on fact verification and multiple-choice tasks. As shown in Figure 5, our training did not reduce the performance of these tasks compared to standard RAG. In fact, in some cases, we even observed a slight improvement. One explanation is that our training strategy encourages collaborative output that leverages the capabilities of the LLM, rather than injecting knowledge directly, and thus possesses a certain degree of universality.

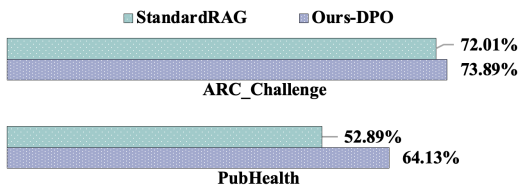

Figure 5: Illustration of accuracy changes when transferring to non-QA tasks, with accuracy as the metric.

4.10 When the Number of K Changes

In order to better explore the robustness of our CoCoA with respect to the number of documents, we set K to vary in the interval [1, 3, 5, 10, 15, 20]. The results are shown in Fig. 6. Overall, our method outperforms StandardRAG across different values of K. Moreover, our method achieves stronger performance than StandardRAG when given less context. We speculate that this is because our model can better utilize internal knowledge, especially when given less information. However, our advantage decreases when the number of documents is too large. We speculate that this is due to the long context bottleneck of the model.

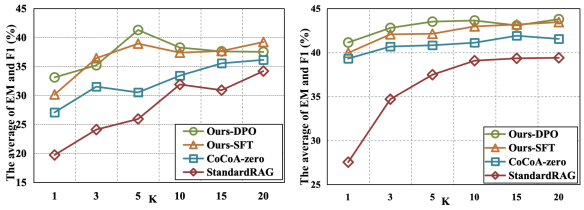

Figure 6: Performance varies with passage num: 2WikiMQA (left), WebQuestions (right).

In summary, our method demonstrates strong robustness across different context sizes and provides a practical solution in settings with limited external information or constrained retrieval capacity.

4.11 Inference Efficiency

To evaluate the prospect of CoCoA in practical applications, we study its inference efficiency. As shown in Table 4, our CoCoA is three times more expensive than common methods (eg., CON), requiring a budget of about 600 output tokens. However, compared to reasoning models (eg., R1), the cost still has a significant advantage, especially when the performance is improved. This achieves a trade-off between performance and efficiency.

Method	Avg. Input	Avg. Output	Avg(EM,F1)
CoT	904.184	195.882	23.83
R1-8B	829.186	818.200	31.30
CoCoA	993.184	609.742	**41.29**

Table 4: Efficiency analysis on 2WikiMQA, with average token count reported.

5 Conclusion

We investigate the challenge of parametric–retrieved knowledge synergy and introduce **CoCoA**, a RAG framework that improves LLM performance. By leveraging a two-stage multi-agent pipeline, CoCoA-zero integrates internal and external knowledge and provides self-synthesized supervisory signals. With long-chain training, CoCoA delivers strong results on QA tasks, demonstrating its effectiveness and offering insights into long-chain reasoning and collaborative agent training for knowledge-intensive applications.

Limitations

While CoCoA has demonstrated excellent performance and provided valuable insights into collaboration with parametric and retrieved knowledge, there are still some limitations:

- The current design focuses on a specific agent collaboration pattern via long-chain training. Its applicability to broader or alternative multi-agent architectures remains to be examined.
- Although the approach performs robustly under limited supervision, its scaling with respect to larger models and datasets has not been systematically explored.
- Although the performance has been improved, the token consumption has increased, which has certain limitations in practical applications. How to accelerate reasoning is still a future research direction.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

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*, pages 1533–1544.

Qinyuan Cheng, Xiaonan Li, Shimin Li, Qin Zhu, Zhangyue Yin, Yunfan Shao, Linyang Li, Tianxiang Sun, Hang Yan, and Xipeng Qiu. 2024. Unified active retrieval for retrieval augmented generation. *arXiv preprint arXiv:2406.12534*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2g: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020a. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020b. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint arXiv:2011.01060*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. 2024a. Sure: Summarizing retrievals using answer candidates for open-domain qa of llms. *arXiv preprint arXiv:2404.13081*.

Kiseung Kim and Jay-Yoon Lee. 2024. Re-rag: Improving open-domain qa performance and interpretability with relevance estimator in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22149–22161.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024b. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2421–2431.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*, 7.

Jinming Nian, Zhiyuan Peng, Qifan Wang, and Yi Fang. 2024. W-rag: Weakly supervised dense retrieval in rag for open-domain question answering. *arXiv preprint arXiv:2408.08444*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6207–6227, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö Arık. 2024a. Astute rag: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. *arXiv preprint arXiv:2410.07176*.

Haoyu Wang, Ruirui Li, Haoming Jiang, Jinjin Tian, Zhengyang Wang, Chen Luo, Xianfeng Tang, Monica Cheng, Tuo Zhao, and Jing Gao. 2024b. Blendfilter: Advancing retrieval-augmented large language models via query generation blending and knowledge filtering. *arXiv preprint arXiv:2402.11129*.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9414–9423.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10303–10315.

Yuhao Wang, Ruiyang Ren, Junyi Li, Wayne Xin Zhao, Jing Liu, and Ji-Rong Wen. 2024c. Rear: A relevance-aware retrieval-augmented framework for open-domain question answering. *arXiv preprint arXiv:2402.17497*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhepei Wei, Wei-Lin Chen, and Yu Meng. 2024. Instructrag: Instructing retrieval-augmented generation via self-synthesized rationales. *arXiv preprint arXiv:2406.13629*.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang. 2025. Improving retrieval augmented language model with self-reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 25534–25542.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Kaixin Ma, Hongwei Wang, and Dong Yu. 2023. Chain-of-note: Enhancing robustness in retrieval-augmented language models. *arXiv preprint arXiv:2311.09210*.

Shengbin Yue, Siyuan Wang, Wei Chen, Xuanjing Huang, and Zhongyu Wei. 2025. Synergistic multi-agent framework with trajectory learning for knowledge-intensive tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25796–25804.

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024. Raft: Adapting language model to domain specific rag. *arXiv preprint arXiv:2403.10131*.

A Dataset

Here, we introduce in detail the datasets we used, which are four datasets on four tasks.

2WikiMultiHopQA (Ho et al., 2020b) and **HotpotQA** (Ho et al., 2020a): Both datasets are multi-hop question answering datasets based on Wikipedia. Considering the limitation of experimental cost, we used the sub-sampling set published by Trivedi et al. (2022); Kim et al. (2024a), which is obtained by extracting 500 questions from the validation set of each dataset.

WebQuestions (Berant et al., 2013): Constructed from questions posed by the Google Suggest API, where the answers are specific entities listed in Freebase.

TriviaQA (Joshi et al., 2017): A compilation of trivia questions paired with answers, both originally pulled from online sources.

Training Data We sampled subsets from the training sets of HotpotQA (Ho et al., 2020a), 2WikiMultiHopQA (Ho et al., 2020b) and WebQuestions (Berant et al., 2013), then used the CoCoA-zero framework to synthesize data and filtered them with gold answers. Finally, we selected 6.8k filtered samples, including 3k, 3k, and 0.8k from the three datasets, respectively. For the DPO training data, we screen out 1151 samples, which are the ones that are answered incorrectly by zero-shot but correctly by the CoCoA-zero. For each sample, we gathered 5 relevant passages using the most common retriever Contriever (Izacard et al., 2021).

B Baselines

We selected several of the most representative methods for comparison.

- **StandardRAG**, which is the most classic "retrieve-then-read" paradigm.

Task Type	Datasets	# Samples
Multi-HopQA	2WikiMultiHopQA	500
	HotpotQA	500
OpenQA	WebQuestions	2032
	TriviaQA	11313

Table 5: Description of tasks and evaluation datasets.

- **Chain-Of-Thought** (Wei et al., 2022): Uses CoT prompting to generate reasoning steps before producing the final answer.
- **Chain-Of-Note** (Yu et al., 2023): Refines the retrieved passages prior to answering.
- **GenRead** (Yu et al., 2022): Generates self-contained intermediate context to answer, effectively replacing retrieval with generation.
- **SURE** (Kim et al., 2024a): Conditional summarization followed by multiple validation.
- **Self-RAG** (Asai et al., 2023): Employs adaptive retrieval and self-reflection to decide when and how to use external knowledge.
- **DeepSeek-R1-Distill-8B** (Guo et al., 2025): A distilled LLaMA-8B model released by DeepSeek-R1, trained on reasoning data.
- **InstructRAG** (Wei et al., 2024): Denoising training using self-synthesized data.

Baseline Setting. We followed the original settings for almost all experiments. All retrieval-based methods use top-5 passages. For baselines requiring training, we directly used their weights. Note that InstructRAG directly generates long rationales, the first half of which consists mostly of analysis and citations of the document, resulting in a non-strictly high EM score and a low F1 score. For a fair comparison, we used Qwen2.5-3B to perform answer segmentation to evaluate.

C Training Details

We fine-tune LLaMA3.1-8B with LoRA (r=16, α=16, dropout=0.05) on a maximum input length of 2048. LoRA is applied to attention projection layers. During SFT, we trained for 5 epochs with a batch size of 1, gradient accumulation of 4, and a learning rate of 3e-5. For DPO, a β value of 0.2 is applied, using a sigmoid loss function, while RPO is configured with an α value of 0.2. The learning

rate was set to 5e-6 and other settings are the same as SFT. During inference, we use the vllm (Kwon et al., 2023) accelerated inference framework, and to ensure repeatability, we set the temperature to 0.0. All experiments are conducted on a single A100 GPU with 80GB or 40GB memory.

C.1 Performance of Different Model Sizes

To verify the performance difference of CoCoA-zero under different model sizes, we conducted experiments on performance changes of different model sizes. As shown in Fig. 7, the larger the LLM, the better the performance of CoCoA-zero, and it far exceeds standardRAG. This shows that larger models better support our collaboration and highlights the importance of internal knowledge in stronger LLMs: the more powerful the LLM, the more it should leverage its internal knowledge for question answering.

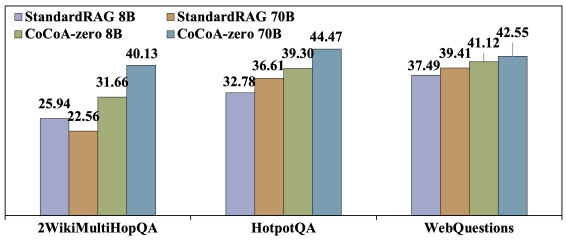

Figure 7: Illustration of performance changes at different model sizes, with Avg(EM,F1) as the metric.

D Performance of Different Retriever

In order to verify the robustness to different retrievers, we selected BM25, DPR, CONTRIEVER and E5 as retrievers. The experimental results are shown in Figure 8. We found that different datasets have different preferences for retrievers, each with its own advantages and disadvantages. Overall, both our CoCoA-zero and CoCoA achieve robust performance on different retrievers.

E Optimization Analysis

We analyze the difference between independent training and long chain training in terms of the form of loss. We simplify the steps in this analysis, i.e., there are only two steps in the chain, pre-generation processing first and then answering.

When the two agents optimize independently, the loss takes the following form:

$$\mathcal{L}_{\text{indep}} = -\log P_\theta\big(s \mid x, d\big) \;-\; \log P_\phi\big(\hat{a} \mid s\big). \tag{9}$$

Here, θ and θ' are optimized independently.

When two agents use long chain optimization, the loss is as follows:

$$\begin{aligned}\mathcal{L}_{\text{chain}} &= -\log P_\theta\big(s, \hat{a} \mid x, d\big) \\ &= -\log P_\theta\big(s \mid x, d\big) \;-\; \log P_\theta\big(\hat{a} \mid x, d, s\big).\end{aligned} \tag{10}$$

Gradient propagation:

The gradient of the first term in Eq. (9) is,

$$\frac{\partial \mathcal{L}_{\text{indep}}}{\partial \theta} = \frac{\partial \left[-\log P_\theta(s \mid x, d)\right]}{\partial \theta} \tag{11}$$

The gradient of the Eq. (10) is,

$$\begin{aligned}\frac{\partial \mathcal{L}_{\text{chain}}}{\partial \theta} &= \frac{\partial \left[-\log P_\theta(s \mid x, d)\right]}{\partial \theta} \\ &\quad + \frac{\partial \left[-\log P_\theta(\hat{a} \mid x, s, d)\right]}{\partial \theta} \\ &= (11) \;+\; \Delta_g\end{aligned} \tag{12}$$

$$\Delta_g \;:=\; \frac{\partial \left[-\log P_\theta(\hat{a} \mid x, s, d)\right]}{\partial \theta}. \tag{13}$$

Here, Δ_g is the additional gradient that the answer-loss naturally back-propagates to the preprocessing parameters when the *same* network θ produces both tokens. In the independent setting $\Delta_g = 0$ by construction, so the preprocessor never "hears" whether the answer is correct, which is not conducive to the consistency of the response. The chain objective restores this missing credit assignment signal, thus performing a special kind of multi-task learning on both stages, optimizing them instead of each in isolation, potentially helping to escape from local optimal solutions.

F Full Results

We supplemented the detailed results of the ablation experiment as shown in Table 6 and Table 7.

G Prompt Templates

All the prompt templates used by our proposed CoCoA are shown in Table 10 and Table 9. And special instructions are added to section 3.9 corresponding to different tasks as shown in Table 8.

H Case Study

We provide detailed case study in Figure 9, Figure 10, and Figure 11.

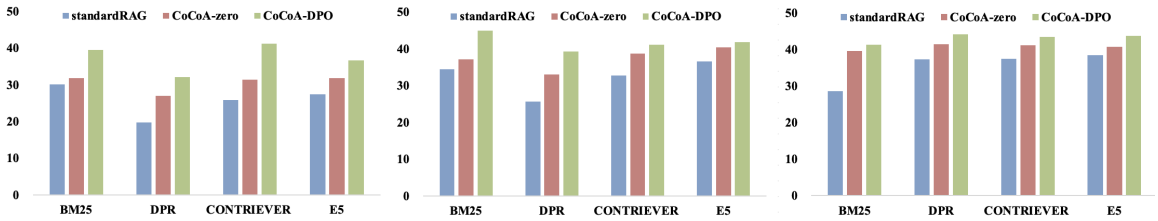

Figure 8: Illustration of performance changes at different retrievers, with Avg(EM,F1) as the metric.

Who lived longer, Aniela Pawlikowska or Erling Jensen?

(Direct Response)

Erling Jensen. ✅

(StandardRAG Response)

Aniela Pawlikowska ❌

(CoCoA Response)

<Internal>
Aniela Pawlikowska was a Polish actress who lived a long and successful career in the film industry. Born in 1907, she was active in the entertainment industry for over four decades, appearing in numerous films and stage productions. Her longevity in the industry suggests that she may have lived a long and healthy life. Erling Jensen, on the other hand, was a Danish actor who was also active in the film industry. While his career was notable, there is limited information available about his personal life and longevity. Without more information, it is difficult to determine which individual lived longer.
<\Internal>
<External>
Aniela Pawlikowska died in 1945, aged 44, while Erling Jensen died in 2015, aged 93. This suggests that Erling Jensen lived longer than Aniela Pawlikowska. (Passage: Anni Jensen Annis "Big Red" Jensen (September 20, 1921 – January 10, 2015) was an American roller derby skater.)
<\External>
<Thinking>
First, let's analyze the information from the internal reasoning path. Aniela Pawlikowska was born in 1907 and was active in the entertainment industry for over four decades, which suggests that she may have lived a long and healthy life. However, there is no specific information about her age at death. Next, let's analyze the information from the external reasoning path. Aniela Pawlikowska died in 1945, aged 44, while Erling Jensen died in 2015, aged 93. This suggests that Erling Jensen lived longer than Aniela Pawlikowska.
<\Thinking>
<Answer>
Erling Jensen. ✅
<\Answer>

Figure 9: Direct generation is correct, while standard RAG generation is incorrect. This is due to external noise causing RAG to misjudge, while CoCoA performs well in synergy.

Method	2WikiMultiHopQA			HotpotQA			WebQuestions		
	EM	F1	Avg	EM	F1	Avg	EM	F1	Avg
CoCoA-zero	31.40	31.92	31.66	37.40	41.20	39.30	43.11	39.13	41.12
-w/o Thinking	30.00	30.76	30.38	36.00	38.34	37.17	39.17	40.32	39.75
-w/o Internal	22.60	23.93	23.26	34.00	39.11	36.56	40.01	38.20	39.10
-w/o External	28.40	29.53	28.97	30.00	31.92	30.96	39.81	38.13	38.97
Zero-Shot	17.60	19.51	18.55	33.20	36.81	35.01	34.45	36.31	35.38
Standard RAG	26.80	25.07	25.94	31.40	34.16	32.78	37.65	37.32	37.49

Table 6: Ablation study of internal/external induction and reasoning in decision making. In addition, a zero-shot method for explicit internal and external knowledge integration is added for comparison. For simplicity and fairness, the average of EM and F1 is used as the metric.

Method	2WikiMultiHopQA			HotpotQA			WebQuestions		
	EM	F1	Avg	EM	F1	Avg	EM	F1	Avg
Long-DPO$_{8B}$	42.00	40.58	41.29	39.00	43.39	41.20	44.83	42.21	43.52
Long-SFT$_{8B}$	41.00	36.87	38.94	39.40	46.31	42.86	42.96	41.32	42.14
Short-SFT$_{8B}$	28.60	28.03	28.31	39.00	42.15	40.58	41.19	38.48	39.84
Short-SFT$_{8B\times3}$	35.00	32.81	33.91	37.60	42.48	40.04	41.29	38.96	40.13

Table 7: Ablation study of the training strategy for CoCoA. For simplicity and fairness, the average of EM and F1 is used as the metric

Task	**Task Instruction**
ARC-C	Given four answer candidates, A, B, C and D, choose the best answer choice. Please answer with the capitalized alphabet only, without adding any extra phrase or period. Do not exceed one word.
PubHealth	Is the following statement correct or not? Say true if it's correct; otherwise say false. Don't capitalize or add periods, just say "true" or "false". Do not exceed one word.

Table 8: Full list of instructions used during zero-shot evaluations. For open-domain QA, we don't use any task specific instruction.

Task:Prompt used by "CoCoA"
Instruction: 1. First, provide background for the question. Write a passage that is relevant to the question only based on your knowledge. 2. Second, refer to the provided passages to generate a summary. Cite and write a passage that is relevant to the question only based on the provided passages. 3. Third, refer to the information from the above two sources, verify the accuracy of the facts and the consistency of the logic, and predict the final answer. ### Passages:\n{$passages$}\n ### Question:\n{$question$} ### Generate Format: <Internal>\nxxx (your background based on your knowledge)\n<\\Internal> <External>\nxxx (your summary based on the provided passages)\n<\\External> <Thinking>\nxxx\n<\\Thinking> <Answer>\nxxx (your short answer consisting of only a few words)<\\Answer>

Table 9: The prompt used by "CoCoA".

Task	Task Instruction
External Candidate	### Passages:\n {$passages$}\n\n ### Instruction:\n Answer the question below concisely in a few words.\n\n ### Input:\n{$question$}\n
External Induction	### Instruction:\n Refer to the provided passages to generate a summary that meets the following conditions:\n 1. Cite and Write a passage that can support the prediction about the question only based on the provided passages.\n 2. No more than 200 words.\n 3. Do not respond with anything other than the S̈ummary:\n ### Passages:\n {$passages$}\n\n ### Question:\n {$question$}\n ### Prediction:\n {$answer$}\n\n ### Generate Format:\n ### Summary: xxx\n
Internal Candidate	### Instruction:\n Answer the question below concisely in a few words.\n\n ### Input:\n{$question$}\n
Internal Induction	### Instruction:\n Please provide background for the question that meets the following conditions:\n 1. Write a passage that can support the prediction about the question only based on your knowledge.\n 2. No more than 200 words.\n 3. Do not respond with anything other than the B̈ackground:\n ### Question:\n {$question$}\n ### Prediction:\n {$answer$}\n\n ### Generate Format:\n ### Background: xxx\n
Decision-Making	### Internal Reasoning Path: \n{$induction_{in}$}\n\n ### Internal Prediction 1: \n{$answer_{in}$}\n\n ### External Reasoning Path: \n{$induction_{ex}$}\n\n ### External Prediction 2: \n{$answer_{ex}$}\n\n ### Instruction:\n Refer to the information from the above two sources, verify the accuracy of the facts and the consistency of the logic, and choose the best prediction. ### Question:\n{$question$}\n ### Generate Format:\n ### Thingking: xxx (Please think step by step)\n ### Short Answer: xxx (just in a few words)\n

Table 10: A list of prompts used by CoCoA-zero.

Which film came out first, Kansas City Confidential or Men O' War?

(Direct Response)

Kansas City Confidential. ❌

(StandardRAG Response)

Men O' War. ✅

(CoCoA Response)

<Internal>
Kansas City Confidential is a 1952 American film noir directed by Phil Karlson, starring John Payne, Preston Foster, and Jack Kelly. The film is a classic example of the film noir genre, known for its dark and gritty portrayal of crime and corruption. Men O' War, on the other hand, is a 1955 American drama film directed by Anthony Mann, starring Jack Palance and Lee J. Cobb. The film is a drama about a group of sailors on a cargo ship, exploring themes of loyalty, duty, and the human condition. Given the release years of these two films, it is likely that Kansas City Confidential came out first, as it was released in 1952, three years before Men O' War.
<\Internal>
<External>
Men O' War is the third sound film starring Laurel and Hardy, released on June 29, 1929. This information supports the prediction that Men O' War came out first, as it was released in 1929, more than 23 years before Kansas City Confidential, which was released in 1952.
<\External>
<Thinking>
First, let's look at the internal reasoning path, which states that Kansas City Confidential was released in 1952, three years before Men O' War, which was released in 1955. This suggests that Kansas City Confidential came out first. However, we need to verify the accuracy of this information.\n\nNext, let's examine the external reasoning path, which provides more specific information about the release date of Men O' War, stating that it was released on June 29, 1929. This information directly contradicts the internal reasoning path, which suggests that Men O' War was released in 1955.
<\Thinking>
<Answer>
Men O' War. ✅
<\Answer>

Figure 10: Direct generation is incorrect, while standard RAG is correct. This is due to internal hallucinations causing misjudgment, while CoCoA performs well.

Who is Margaret Of Valois, Countess Of Blois's maternal grandfather?

(Direct Response)

Henry IV of France. ❌

(StandardRAG Response)

Nicola Orsini ❌

(CoCoA Response)

<Internal>
Margaret of Valois, Countess of Blois, was born in 1553 as the daughter of Henry II of France and Catherine de' Medici. Her father, Henry II, was the King of France from 1547 until his death in 1559. Catherine de' Medici, Margaret's mother, was an Italian noblewoman and the daughter of Lorenzo II de' Medici, Duke of Urbino, and Madeleine de La Tour d'Auvergne. Catherine de' Medici was a member of the powerful Medici family, which had a significant influence on the politics of Italy and Europe. As the daughter of Lorenzo II de' Medici, Catherine's father was a member of the Medici family, and her family's influence would have a lasting impact on Margaret's life and upbringing.
<\Internal>
<External>
Margaret of Valois, Countess of Blois's maternal grandfather was Charles II of Naples. This can be supported by the passage: \"Her maternal grandparents were Charles II of Naples and Maria Arpad of Hungary.\" (Passage 2: Joan of Valois, Countess of Hainaut)
<\External>
<Thinking>
To determine Margaret of Valois, Countess of Blois's maternal grandfather, we need to analyze the information provided in both sources. The first source mentions that Catherine de' Medici, Margaret's mother, was the daughter of Lorenzo II de' Medici, Duke of Urbino. The second source mentions that Margaret of Valois, Countess of Blois is not the same person as the one in the first source, but rather Joan of Valois, Countess of Hainaut. However, the second source also mentions that Joan's maternal grandparents were Charles II of Naples and Maria Arpad of Hungary.
<\Thinking>
<Answer>
Charles II of Naples ✅
<\Answer>

Figure 11: Direct generation is incorrect, and standard RAG generation is also incorrect. CoCoA synergizes well, benefiting from the interaction between knowledge induction and high-level decision-making to stimulate model capabilities.