Chain-of-Note: Enhancing Robustness in Retrieval-Augmented Language Models

Announcement Date	@2024년 12월 19일
Conference Name	EMNLP 2024
Keywords	CoT LLM RAG

Abstract

- Retrieval-augmented language model (RALM) 은 외부 지식 소스를 활용하여 factual hallucination (사실적 환각)을 완화하는데 중요한 진전을 보여줌
- 그러나, 검색된 정보의 신뢰성이 항상 보장되지 않으며, 관련 없는 데이터를 검색하는 것은 응답 생성을 혼란스럽게 할 수 있음
- 더불어, standard RALM은 종종 검색된 정보의 간섭으로 인해 고유 지식(intrinsic knowledge)을 무시하는 경우가 많음
- 검색된 정보가 관련이 없는 경우에는, RALM은 1. 고유 지식을 활용해야 하며 2. 고유 지식과 검색된 지식 모두 없는 경우 “모름” 으로 응답하여 hallucination을 피해야 함

- 그래서 본 논문에서는, 노이즈가 있는 관련 없는 문서와 알 수 없는 상황에 직면하는 RALM의 강인성을 향상시키기 위한 새로운 접근 방식인 Chain - of - Note (CoN) 을 제안
- CoN의 핵심 아이디어는 각 검색된 문서에 대한 순차적 읽기 노트를 생성하여, 주어진 문서에 대한 관련성을 철저하게 평가하고 이 정보를 통합하여 최종 답변을 생성하는 것
- 실험 결과에 따르면, CoN을 적용한 GPT-4는 chain-of-thought 접근방식을 능가함
- 또한, 저자는 GPT-4 를 활용하여 10K개의 CoN data를 생성하고, 이후 LLaMa-2 7B 모델에 학습
- 네 개의 open-domain QA 벤치마크에서의 실험 결과는 CoN이 적용된 fine-tuned RALM이 standard fine-tuned RALM보다 유의미하게 성능이 좋았음

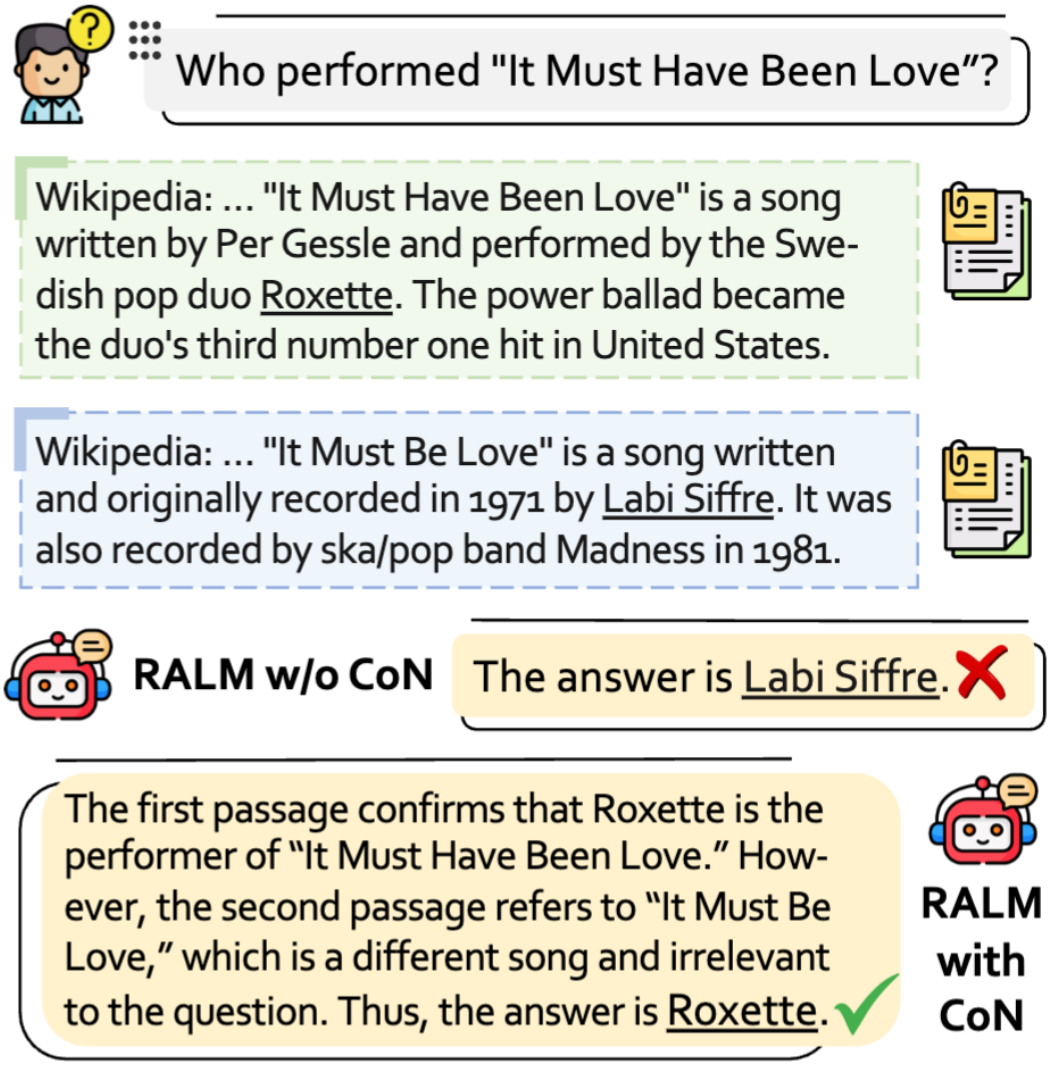


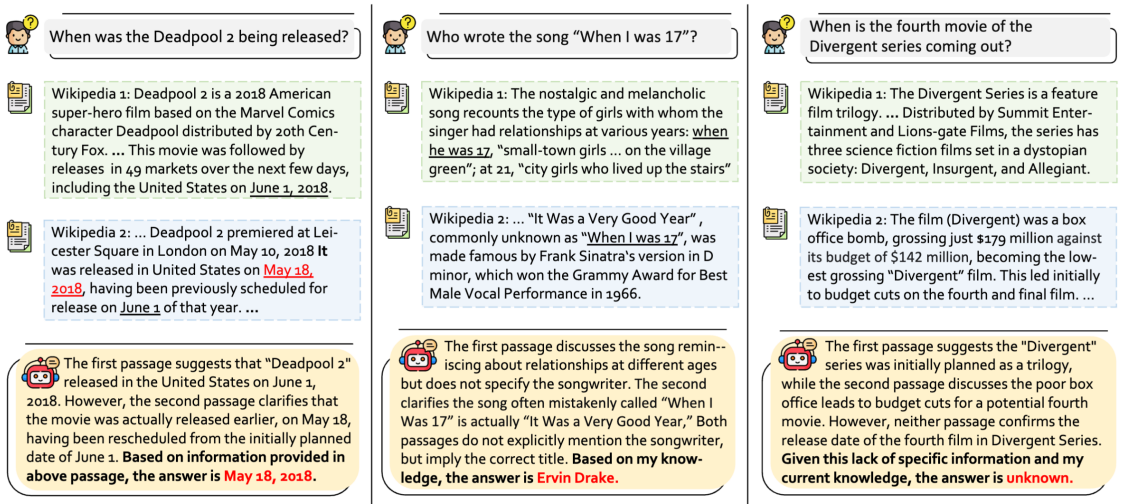
Figure 1: Compared with the current RALMs, the core idea behind CHAIN-OF-NOTE (CON) is to generate sequential reading notes for the retrieved documents, ensuring a systematic assessment of their relevance to the input question before formulating a final response.

Proposed Method

Overview

- CHAIN - OF - NOTE (CoN) 를 제안
- CoN framework는 검색된 문서를 위한 sequential reading notes를 생성하여 외부 문서에서 검색된 정보의 관련성과 정확성을 체계적으로 평가할 수 있음
- Sequential Reading Notes를 생성함으로써 모델은 각 문서의 쿼리 적합성을 평가할 뿐만 아니라 이러한 문서 내에서 가장 중요하고 신뢰할 수 있는 정보 조각(reliable pieces) 을 식별할 수 있음
- 이러한 과정은 불필요하거나 신뢰성이 낮은 내용을 걸러내어 보다 정확하고 맥락적으로 관련 있는 응답을 이끌어냄

The CHAIN-OF-NOTE Framework



(a) CoN: Relevant -> Find the answer (b) CoN: Irrelevant -> Infer the answer (c) CoN: Irrelevant -> Answer *Unknown*

Figure 2: Illustration of the CHAIN-OF-NOTE (CON) framework with three distinct types of reading notes. Type (a) depicts the scenario where the language model identifies a document that directly answers the query, leading to a final answer formulated from the retrieved information. Type (b) represents situations where the retrieved document, while not directly answering the query, provides contextual insights, enabling the language model to integrate this context with its inherent knowledge to deduce an answer. Type (c) illustrates instances where the language model encounters irrelevant documents and lacks the necessary knowledge to respond, resulting in an "unknown" answer. This figure exemplifies the CoN framework's capability to adaptively process information, balancing direct information retrieval, contextual inference, and the recognition of its knowledge boundaries.

- CoN framework 는 세 가지 유형으로 나누어짐

1. 유형 (a)는 언어 모델이 쿼리에 직접 답변하는 문서를 식별하여, 검색된 정보를 기반으로 최종 답변을 생성하는 시나리오를 나타냄

2. 유형 (b)는 검색된 문서가 쿼리에 직접 답변하지는 않지만, 문서가 제공하는 맥락적 통찰을 통해 언어 모델이 이 맥락을 본질적인 지식(inherent knowledge) 과 통합하여 답변을 도출하는 상황
3. 유형 (c)는 언어 모델이 관련 없는 문서를 만나거나 필요한 지식을 갖추지 못해 “알 수 없음(unknown)” 이라는 답변을 도출하는 상황

Data Collection and Model Training

- 모델이 이러한 reading notes 를 생성할 수 있도록 하려면 적합한 training data 수집이 필요
- 각 reading notes 에 대한 수동 주석 작업은 자원이 많이 소모되므로, 저자는 GPT-4 를 활용하여 노트 데이터를 생성
- NQ Training Dataset 에서 10,000개의 질문을 랜덤 샘플링하여 이 과정을 시작
    - NQ : Natural questions: A benchmark for question answering research (2019 TACL)
- GPT-4에게 구체적인 instruction과 3가지 독서 노트 유형에 대한 컨텍스트 예제 (figure 2) 를 제공

- GPT-4 를 통해 10,000개의 훈련 데이터를 수집한 후, LLaMa-2 7B 모델을 훈련하여 Chain-Of-Note 출력을 생성할 수 있는가? 를 검증
- 내부 모델은 각 문서에 대한 reading notes 를 순차적으로 생성하여 입력 쿼리에 대한 관련성을 평가하는 방법을 학습
- 응답은 문서의 관련성을 바탕으로 생성되어 정확도를 향상시키고 잘못된 정보를 줄임

Hybrid Training for Better Efficiency

- CON 의 생성은 추론 비용 (inference cost)를 증가시켜 실제 사용에 방해가 될 수 있음
- 이를 해결하기 위해 Hybrid Training 을 실험
- 훈련 시간의 50%를 standard RALM에 할당하고, 나머지 50%를 CoN이 포함된 RALM에 할당
- 이 전략을 통해 모델이 훈련 중 중간 추론 단계를 내제화할 수 있도록 함

- 추론 단계에서는 standard RALM prompt만 사용하여 모델이 explicit Reading Notes에 의존하지 않고 답변을 출력하도록 유도
- 이 접근법은 훈련 중 개발된 hidden states 를 활용하여 암묵적인 CoN 추론을 수행하게 함

- 결론적으로, 하이브리드 훈련 전략으로 훈련된 모델은 동일한 추론 시간을 유지하면서도 CoN을 포함했을 때 약간 낮은 성능을 보이는 결과를 얻음

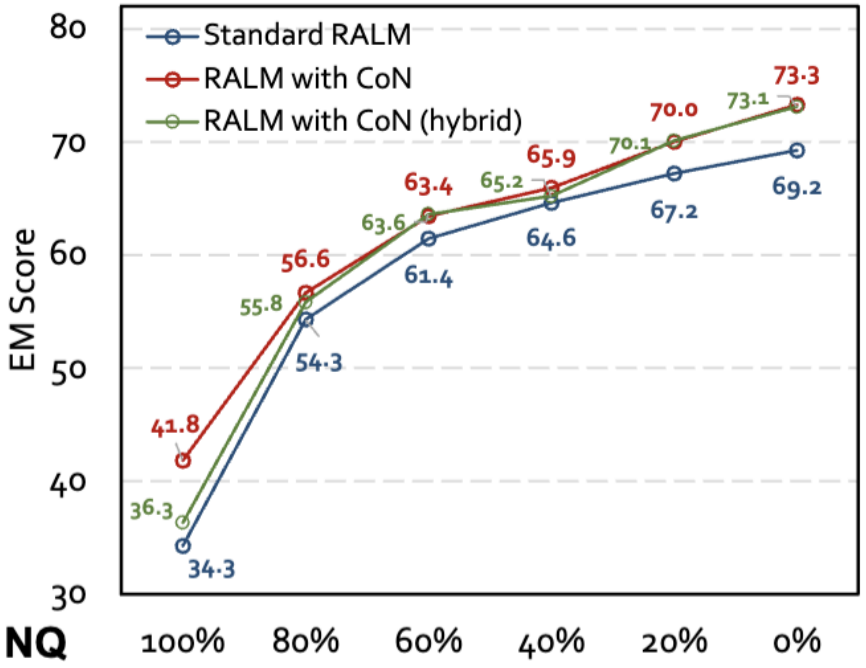


Figure 3: Using a hybrid training strategy demonstrates slightly lower robustness across various noise ratios but consistently better performance than standard RALMs.

Models ↓	Inference Time(s)
Retrieve-Read	0.6104
+ CHAIN-OF-NOTE	12.0192
+ CHAIN-OF-NOTE (hybrid)	0.6074

Table 5: The inference time comparison shows the average decoding time per example on 8×A100 GPUs.

Experiments

- NQ / TriviaQA / WebQ 세 가지 벤치마크 데이터셋을 사용하여 실험 수행
- 또한, 없는 정보에 대해 “Unknown” 이라고 출력하는 강인성 평가를 위해 Real-TimeQA 데이터셋을 사용하여 실험

Models	NQ		TriviaQA		WebQ		Average	
	EM	F1	EM	F1	EM	F1	EM	F1
Backbone language model: LLaMa-2 7B								
QA fine-tune w/o IR	28.80	37.53	63.19	68.61	28.30	42.77	35.98	44.27
SAIL (Luo et al., 2023)*	36.20	44.23	73.20	80.92	27.92	40.65	45.77	55.27
Retrieve-Read (Shi et al., 2023c)	47.39	55.81	74.92	81.53	29.58	43.51	48.49	56.97
+ CHAIN-OF-NOTE (ours)	48.92	57.53	76.27	82.25	32.33	46.68	50.46	58.78
	(+1.53)	**(+1.72)**	**(+1.35)**	**(+0.72)**	**(+2.75)**	**(+3.17)**	**(+1.97)**	**(+1.81)**
Backbone language model: `GPT-4-1106` †								
QA prompt w/o IR	54.0		74.2		56.2		61.5	
Retrieve-Read (Shi et al., 2023c)	61.8		70.6		56.8		63.1	
+ CHAIN-OF-THOUGHT	63.6		71.2		58.4		64.4	
+ CHAIN-OF-NOTE (OURS)	63.8		74.6		58.8		65.7	
	(+2.0)		**(+4.0)**		**(+2.0)**		**(+2.6)**	

Table 2: The RALM, when equipped with CHAIN-OF-NOTE (CON), demonstrates a marginal improvement over the standard RALM in full test set evaluations. Significantly, it outperforms the standard RALM system in scenarios with noisy documents, suggesting that CON can substantially enhance the model’s noise robustness.
* SAIL was designed for retrieval-augmented instruction tuning, and as such, may not be ideally factual QA.
† Evaluating GPT-4 outputs with EM score is challenging; we opt for Accuracy, with reasons outlined in §3.1.3.

1. QA fine-tuned w/o IR : 외부 검색된 정보에 의존 x, 입력 질문에서 직접 답변을 생성하도록 훈련됨 (f : x →y)
2. Retrieve-Read : 질문뿐만 아니라 검색된 문서도 통합하여 답변을 생성하도록 훈련됨
3. Retrieve-Read with CoN : 최종 답변을 형성하기 전에 각 검색된 문서에 대한 reading note 를 생성하도록 훈련됨

*** SAIL 은 Retrieval-Augmented instruction tuning을 위해 설계되었으므로, factual QA에는 적합하지 않을 수 있음

*** Evaluation : EM (Exact Match), F1 Score

Models	Noise Ratio	NQ		TriviaQA		WebQ		Average	
		EM	F1	EM	F1	EM	F1	EM	F1
Retrieve-Read	100%	34.28	41.74	55.30	61.67	29.58	46.34	39.72	49.92
+ CHAIN-OF-NOTE		41.83	49.58	64.30	70.00	36.85	53.07	47.66	57.55
		(+7.55)	(+7.84)	(+9.00)	(+8.33)	(+7.27)	(+6.73)	(+7.94)	(+7.63)
Retrieve-Read	80%	54.28	61.03	73.83	80.02	35.46	52.70	54.52	64.58
+ CHAIN-OF-NOTE		56.63	63.23	75.89	81.24	40.60	56.54	57.70	67.00
		(+2.35)	(+2.20)	(+2.06)	(+1.22)	(+5.14)	(+3.84)	(+3.18)	(+2.42)
Retrieve-Read	60%	61.44	67.94	78.44	83.65	37.01	54.16	58.96	68.58
+ CHAIN-OF-NOTE		63.43	69.33	78.79	84.07	41.26	56.91	61.16	70.10
		(+1.99)	(+1.39)	(+0.35)	(+0.42)	(+4.25)	(+2.75)	(+2.20)	(+1.52)
Retrieve-Read	40%	64.62	71.12	80.56	86.76	38.40	55.60	61.19	71.16
+ CHAIN-OF-NOTE		65.91	72.22	81.72	87.11	42.16	58.15	63.26	72.49
		(+1.29)	(+1.10)	(+1.16)	(+0.35)	(+3.76)	(+2.55)	(+2.07)	(+1.33)
Retrieve-Read	20%	67.21	73.69	81.73	87.89	39.95	56.66	62.96	72.75
+ CHAIN-OF-NOTE		70.00	76.08	82.86	88.24	44.36	60.13	65.74	74.82
		(+2.79)	(+2.39)	(+1.13)	(+0.35)	(+4.41)	(+3.47)	(+2.78)	(+2.07)
Retrieve-Read	0%	69.23	75.57	83.34	89.44	42.24	58.59	64.93	74.53
+ CHAIN-OF-NOTE		73.28	79.86	83.52	88.94	46.16	62.38	67.65	77.06
		(+4.05)	(+4.29)	(+0.18)	(-0.50)	(+3.92)	(+3.79)	(+2.72)	(+2.53)

Table 3: Evaluation on Noise Robustness. The backbone language model is LLaMa-2 7B. The CHAIN-OF-NOTE framework shows superior performance compared to the standard RALM system, particularly notable at higher noise ratios.We explain how we synthesize data with different noise ratios under real-world scenarios in § 3.1.1.

- Noise Ratio 에 따른 2가지 모델의 성능 평가
- 대체적으로 Noise Ratio 가 클수록 CoN 의 효과가 좋은 것을 확인

Models ↓	RealTimeQA		
	EM	F1	RR
Retrieve-Read (Shi et al., 2023c)	15.6	19.9	6.1
+ CHAIN-OF-NOTE (ours)	15.7	20.3	13.0

Table 4: Evaluation on Unknown Robustness. The CON shows better performance than standard RALM system.

- 아까 언급했던 “Unknown”을 답해야 하는 강인성 평가에 쓰이는 RealTimeQA 도 평가
- RR 은 Reject Rate (답이 없어서, 답변을 거부한 수)
- LLaMa-2 의 pre-trained 데이터에 없는 실시간 정보들이 RealTimeQA에 포함되어 있음

- CoN을 적용한 모델이 더 좋은 강인성을 가짐, 특히 RR 에서 유의미한 향상