동계 세미나

Adaptive-RAG와 Query Complexity Classifier 개선 방안: Effectiveness한 정보 검색과 Efficiency 향상

정지민

Research Object for Classifier Design

- Adaptive Retrieval 방식을 이용한 논문에서, Query의 Complexity에 따라 Classifier를 이용해 검색 횟수를 조정하는 아이디어에 흥미를 느껴 이를 기반으로 Classifier 방법론을 공부하고 설계하는 것을 최종 목표로 설정
- 다양한 Classifier 설계 방법론을 학습하는 중이며, Effectiveness를 최대한 유지하며 Efficiency를 향상시키는 것이 핵심 목표

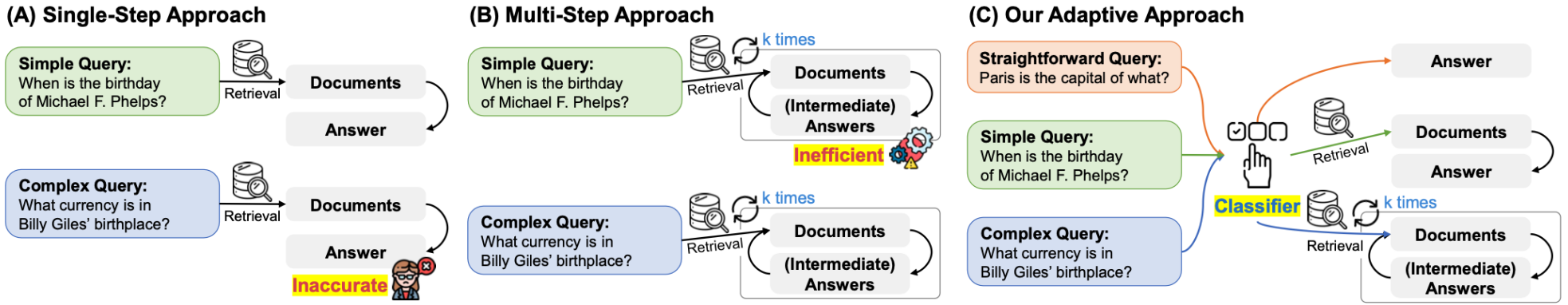
- **Query Complexity Classifier를 다루는 기존 논문에서 Multi-Retrieval을 처리하는 대표적인 방식으로 Query Decomposition과 IRCoT 방식을 확인하여 Query Decomposition 방식의 대표적인 사례인 Layered Query Retrieval과 IRCoT 방식을 적용한 Adaptive-RAG 두 논문의 리뷰를 통해 기존 접근 방식의 한계를 분석하고, 더 나은 Classifier 설계를 위한 인사이트 도출 예정**

Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity

Soyeong Jeong[1] **Jinheon Baek**[2] **Sukmin Cho**[1] **Sung Ju Hwang**[1,2] **Jong C. Park**[1*]
School of Computing[1] Graduate School of AI[2]
Korea Advanced Institute of Science and Technology[1,2]
{starsuzi,jinheon.baek,nelllpic,sjhwang82,jongpark}@kaist.ac.kr

NAACL 2024

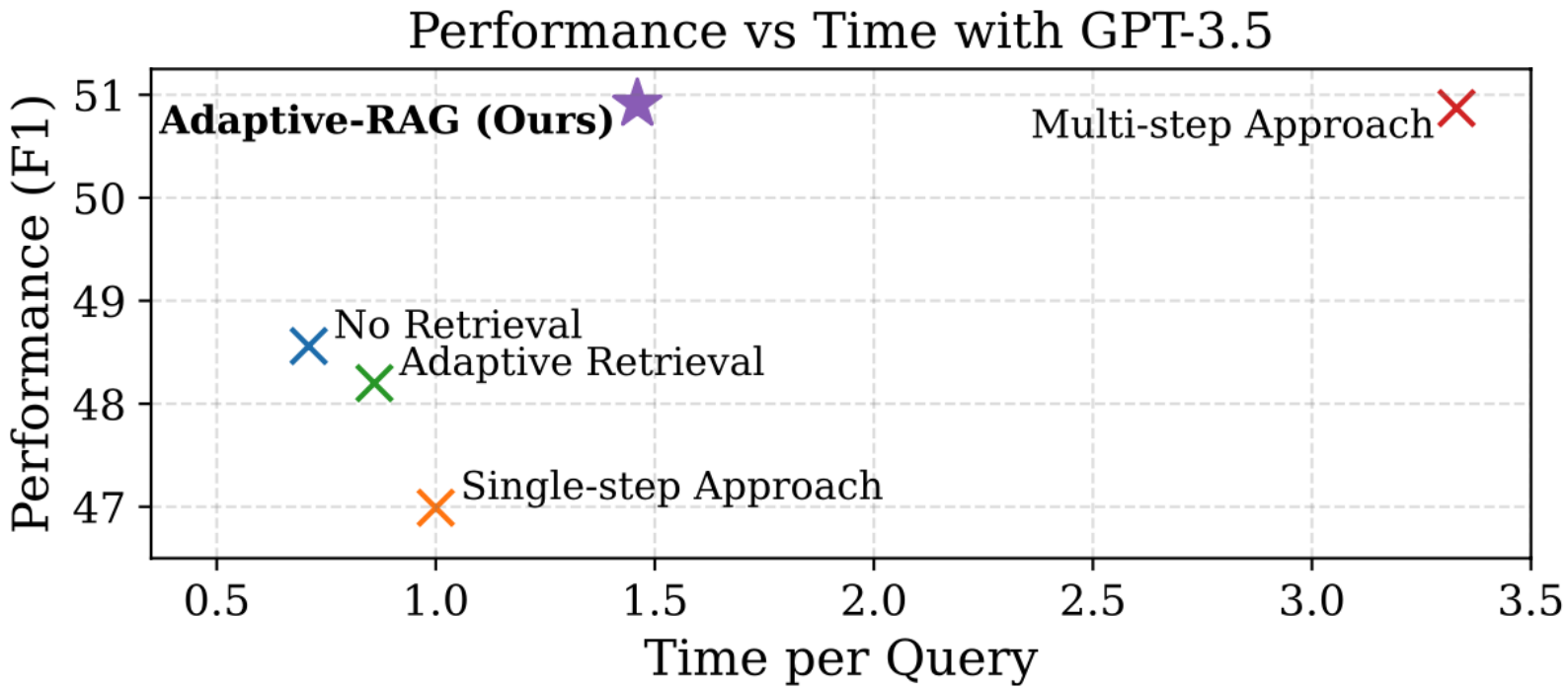
1. Background



- LLM은 parametric memory에만 의존하여 사실적으로 부정확한 답변을 생성
- RAG는 검색을 통해 확보한 non-parametric한 지식을 이용해 LLM을 증강시켜 더 정확한 최신의 정보를 제공 가능
- 기존의 논문은 쿼리의 복잡성이나 필요한 정보 수준을 고려하지 않고, 모든 쿼리에 일관된 Retrieval 전략을 적용하거나 No-Retrieval 또는 Retrieval과 같이 이분법적으로 전략을 선택
- **Adaptive-RAG는 질문 복잡도 Classifier를 통해 주어진 질문의 복잡도를 결정하고, No-Retrieval, Single-Retrieval, Multi-Retrieval 중 가장 적합한 전략을 선택하여 정확하고 효율적으로 검색을 수행**

2. Contribution

" 기존 Adaptive Retrieval 방식의 Classifier는 검색 여부만 판단한 반면, Adaptive-RAG의 Classifier는 질의의 복잡도에 따라 No-Retrieval, Single-Retrieval, Multi-Retrieval 등 적합한 검색 전략을 동적으로 선택할 수 있도록 설계"



3. Training Strategy

- **문제 정의**
- Query에 대해 Complexity를 정확히 예측하는 classifier 학습(FLAN-T5 모델)이 필요
- Query-Complexity 쌍에 대한 Annotation 데이터가 없으므로 자동으로 데이터셋을 구축

- **데이터셋 구축 전략**
- **Step 1: 검색 증강 LLM(FLAN-T5 XL, XXL, GPT-3.5) 결과 기반 Labeling**(6 종류 데이터셋에서 Train Dataset 400개 샘플링)
  - 'A' : 검색 없이(No-Retrieval) 올바른 답변 생성 시
  - 'B' : 단일 단계(Single-Retrieval)으로 올바른 답변 생성 시
  - 'C' : 다단계(Multi-Retrieval)으로 올바른 답변 생성 시
- **Step 2: 벤치마크 데이터셋 기반 보완(Inductive-Bias)**
  - 'B' : Single-Hop Dataset
  - 'C' : Multi-Hop Dataset

Single-Hop Dataset	Multi-Hop Dataset
SQuAD v1.1	MuSiQue
Natural Questions	HotpotQA
TriviaQA	2WikiMultiHopQA

- **학습 및 추론**
- Single-Retrieval 처리: Complexity가 'B'로 분류된 경우, 단일 문서를 Retrieval하여 적합한 결과 선택
- Multi-Retrieval 처리: Complexity가 'C'로 분류된 경우, IRCoT 방식으로 LLM이 추론 단계를 단계적으로 진행하며, 각 단계에서 수행된 Retrieval 결과를 통합하여 최종 답변 생성

4. Evaluation Results

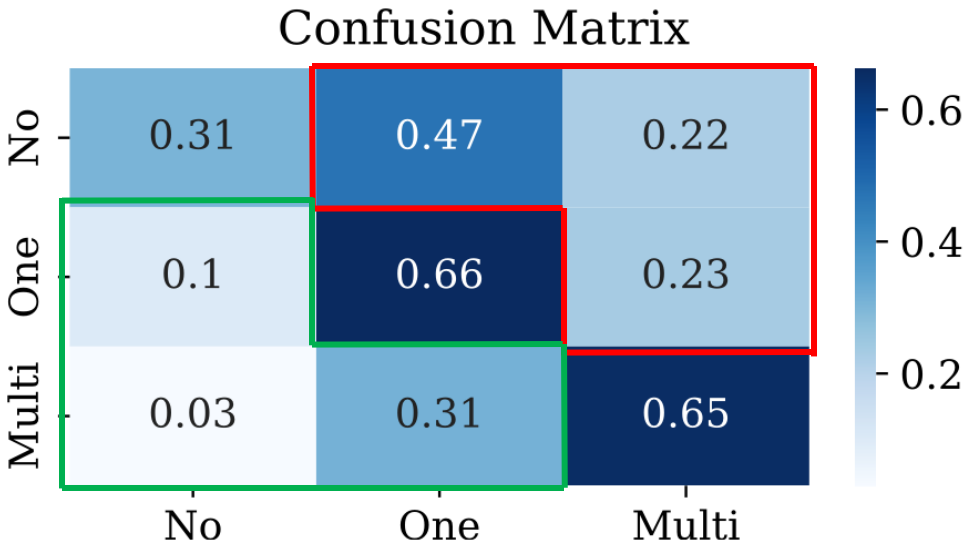
Types	Methods	FLAN-T5-XL (3B) EM	F1	Acc	Step	Time	FLAN-T5-XXL (11B) EM	F1	Acc	Step	Time	GPT-3.5 (Turbo) EM	F1	Acc	Step	Time
Simple	No Retrieval	14.87	21.12	15.97	0.00	0.11	17.83	25.14	19.33	0.00	0.08	35.77	48.56	44.27	0.00	0.71
	Single-step Approach	34.83	44.31	38.87	1.00	1.00	37.87	47.63	41.90	1.00	1.00	34.73	46.99	45.27	1.00	1.00
Adaptive	Adaptive Retrieval	23.87	32.24	26.73	0.50	0.56	26.93	35.67	29.73	0.50	0.54	35.90	48.20	45.30	0.50	0.86
	Self-RAG*	9.90	20.79	31.57	0.72	0.43	10.87	22.98	34.13	0.74	0.23	10.87	22.98	34.13	0.74	1.50
	Adaptive-RAG (Ours)	**37.17**	**46.94**	**42.10**	**2.17**	**3.60**	**38.90**	**48.62**	**43.77**	**1.35**	**2.00**	**37.97**	**50.91**	**48.97**	**1.03**	**1.46**
Complex	Multi-step Approach	39.00	48.85	43.70	4.69	8.81	40.13	50.09	45.20	2.13	3.80	38.13	50.87	49.70	2.81	3.33
Oracle	Adaptive-RAG w/ Oracle	45.00	56.28	49.90	1.28	2.11	47.17	58.60	52.20	0.84	1.10	47.70	62.80	58.57	0.50	1.03

- **평가 지표**
- **effectiveness:** EM(Exact Match), F1, Acc(Accuracy)
- **efficiency:** # of Retrieval-and-Generate Steps, Average Time for Answering Each Query

- **결과 분석**
- Simple Type에 비해 Effectiveness는 높지만 높은 Computational Cost이 발생
- Complex Type에 비해 Effectiveness는 낮지만 Cost Efficient
- Adaptive Type에서 Effectiveness는 제일 높음
- **가장 이상적인 Classifier인 Oracle Classifier를 사용한 경우 Effectiveness, Efficiency 모두 가장 좋은 Performance를 기록**

5. Limitations(efficiency)

Labels	Time/Query (Sec.)	Percentage (%)
No (A)	0.35	8.60
One (B)	3.08	53.33
Multi (C)	27.18	38.07



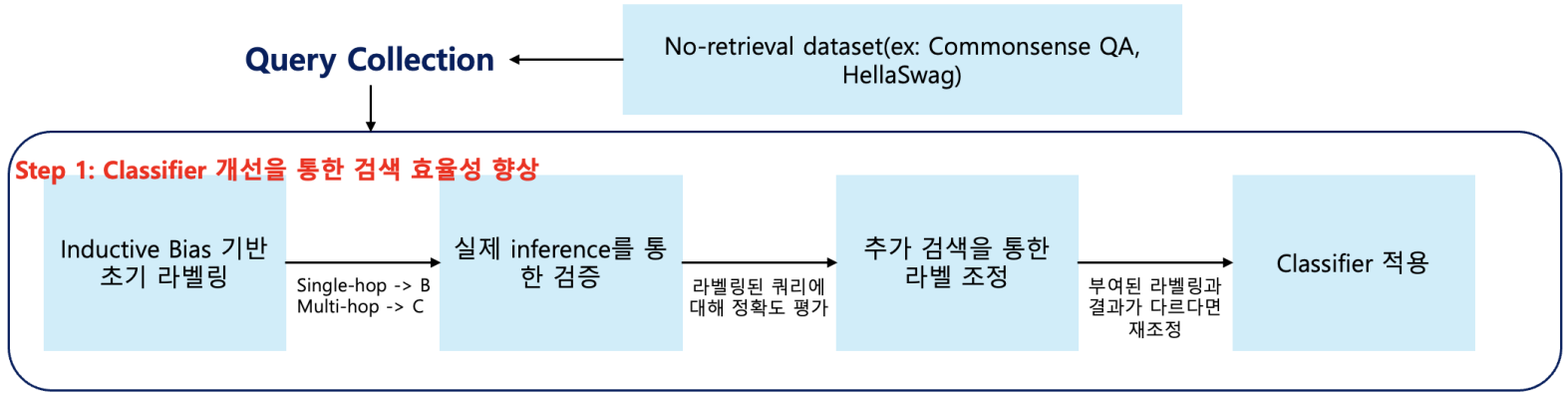
- **Classifier Accuracy 관점: QA Effectiveness는 높지만 Classifier의 낮은 Accuracy로 인한 QA 시스템 Efficiency 저하**
  - Adaptive-RAG는 Classifier에 전적으로 의존하나, Classifier의 Performance가 미흡하여 오분류가 빈번하게 발생
  - 'C'가 'A'로 오분류되는 비율: **3%**, 'B'로 오분류되는 비율: **31%**
  - 'B'가 'A'로 오분류되는 비율: **1%**, 'C'로 오분류되는 비율 : **23%**
  - 'A'가 'B'로 오분류되는 비율: **47%**, 'C'로 오분류되는 비율: **22%**
  - 이러한 오분류는 Retrieval 과정에서 불필요한 단계를 추가하여 QA Efficiency 저하를 야기

- **Dataset 관점: 자동 생성된 라벨 및 Query 데이터의 다양성 부족**
  - 학습 데이터를 수작업으로 라벨링하지 않고, 모델의 예측 결과와 데이터셋의 구조적 특징(Inductive Bias)을 기반으로 라벨링 진행
  - Train 과정에서 Single-Hop, Multi-Hop 두 종류의 데이터셋만 사용하였으며, No-Retrieval Query는 모델의 예측에서 누락되는 경우 라벨링 불가

	QA		Classifier (Accuracy)			
Training Strategies	F1	Step	All	No	One	Multi
Adaptive-RAG (Ours)	**46.94**	**1084**	**54.52**	**30.52**	**66.28**	**65.45**
w/o Binary	43.43	640	60.30	62.19	65.70	39.55
w/o Silver	48.79	1464	40.00	0.00	53.98	75.91

6. Experiment for Understanding Adaptive-RAG

Adaptive-RAG 논문 개선 아이디어



Training Strategies	QA		Classifier (Accuracy)			
	F1	Step	All	No	One	Multi
Adaptive-RAG (Ours)	**46.94**	**1084**	**54.52**	**30.52**	**66.28**	**65.45**
w/o Binary	43.43	640	60.30	62.19	65.70	39.55
w/o Silver	48.79	1464	40.00	0.00	53.98	75.91

Single-Hop Dataset	Multi-Hop Dataset
SQuAD v1.1	MuSiQue
Natural Questions	HotpotQA
TriviaQA	2WikiMultiHopQA

- **Classifier의 Accuracy 향상에 초점**

- Silver Data만 사용하는 경우 많은 검색 단계에서 누적되는 Irrelevant Documents나 Noise로 인해 Multi-Retrieval의 Classifier Accuracy가 감소한다. -> 이후 논문에서 소개 할 Relevance Classifier 도입 필요
- Multi-Retrieval의 Classifier의 Accuracy를 높이기 위해 1차적으로 Inductive Bias 기반 라벨링을 수행하여, 자동 라벨링 대비 Reliability와 Accuracy 향상(초기 단계에서 고품질 라벨을 제공함으로써 Noise 감소)
- Inductive Bias를 활용한 초기 라벨링 후 Inference 과정을 통해 불필요한 검색 횟수를 줄이고, 올바른 라벨에 집중함으로써 Noise 및 Inference 오류 확률 감소
- No-Retrieval Query의 분류 정확도 향상을 위해 No-Retrieval Query 데이터셋 추가로 도입

7. Limitations of Experiment

- **Inductive Bias 기반 라벨링 정확도 부족**
• 762개 데이터셋으로 실험해 본 결과 Inductive Bias와 LM의 동일 라벨링 비율은 31.76%(242개)에 불과
• 68.24%는 다른 라벨링 결과를 보여 개선 효과가 제한적
> Inductive Bias 라벨링으로 Reliability 향상 얻기 어려움

- **검색 전략의 구조적 한계**
• B로 초기 라벨링된 경우: 검색 성공 시 A로도 검색이 가능한지 추가 Validation 필요
• C로 초기 라벨링된 경우: 검색 성공 시에도 A와 B로 검색이 가능한지 추가 Validation 필요
> 검색 횟수를 줄이는 데 한계 존재

- **연구 방향 재설정**
• Layered-Query-Retrieval 논문을 리뷰하며 Adaptive-RAG 개선 아이디어를 재검토
• Layered-Query-Retrieval의 Classifier를 분석해 더 나은 Classifier 아이디어 제안

- **Layered Query Retrieval 논문의 배경**
• 목적: Adaptive-RAG의 효율성과 효과성을 개선하기 위해 제안된 논문
• 특징: Adaptive-RAG와 유사한 실험 세팅

Article

Layered Query Retrieval: An Adaptive Framework for Retrieval-Augmented Generation in Complex Question Answering for Large Language Models

Jie Huang [1,2], Mo Wang [1,2,*], Yunpeng Cui [1,2], Juan Liu [1,2], Li Chen [1,2], Ting Wang [1,2], Huan Li [1,2] and Jinming Wu [1,2]

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[2] Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081, China
* Correspondence: wangmo@caas.cn

Applied Sciences 14(23), 2024

1. Evaluation Results

Type	Method	SQuAD				Natural Questions				TriviaQA			
		F1	Acc	Step	Time	F1	Acc	Step	Time	F1	Acc	Step	Time
Simple	Non-Retrieval	10.50	5.00	0.00	0.11	19.00	15.60	0.00	0.13	31.80	27.00	0.00	0.13
	Single-Step Retrieval	39.30	34.00	1.00	1.00	47.30	44.60	1.00	1.00	62.40	60.20	1.00	1.00
Complex	Multi-Step Retrieval	35.60	29.60	4.52	9.03	47.80	44.20	5.04	10.18	62.40	60.20	5.28	9.22
	Adaptive Retrieval	23.10	17.60	0.50	0.55	36.00	33.00	0.50	0.56	46.90	42.60	0.50	0.56
Adaptive	Self-RAG	11.20	18.40	0.63	0.50	39.00	33.60	0.63	0.17	29.30	57.00	0.68	0.45
	Adaptive-RAG	38.30	33.00	1.37	2.02	47.30	44.60	1.00	1.00	60.70	58.20	1.23	1.54
	Our Method	**46.50**	**43.40**	3.73	4.30	**60.16**	**58.8**	3.24	3.41	**70.20**	**68.80**	3.84	4.50

Type	Method	MuSiQue				HotpotQA				2Wiki			
		F1	Acc	Step	Time	F1	Acc	Step	Time	F1	Acc	Step	Time
Simple	Non-Retrieval	10.70	3.20	0.00	0.11	22.71	17.20	0.00	0.11	32.04	27.80	0.00	0.10
	Single-Step Retrieval	22.80	15.20	1.00	1.00	46.15	36.40	1.00	1.00	47.90	42.80	1.00	1.00
Complex	Multi-Step Retrieval	31.90	25.80	3.60	7.58	56.54	47.00	5.53	9.38	58.85	55.40	4.17	7.37
	Adaptive Retrieval	15.80	8.00	0.50	0.55	32.22	25.00	0.50	0.55	39.44	34.20	0.50	0.55
Adaptive	Self-RAG	8.10	12.00	0.73	0.51	17.53	29.60	0.73	0.45	19.59	38.80	0.93	0.49
	Adaptive-RAG	31.80	26.00	3.22	6.61	53.82	44.40	3.55	5.99	49.75	46.40	2.63	4.68
	Our Method	**41.97**	**26.55**	3.88	4.45	**69.96**	**53.8**	3.42	3.92	**54.65**	**37.60**	3.88	4.32

Single-Hop Dataset	Multi-Hop Dataset
SQuAD v1.1	MuSiQue
Natural Questions	HotpotQA
TriviaQA	2WikiMultiHopQA

- Multi-Hop 데이터셋의 경우 Adaptive-RAG에 비해 대체로 높은 F1, Accuracy와 개선된 Time을 기록
- Single-Hop 데이터셋의 경우 Adaptive-RAG에 비해 향상된 F1, Accuracy를 기록하였지만, **Time 관점에서 Multi-Hop 데이터셋과 유사하고, Adaptive-RAG에 비해 악화**

2. Workflow

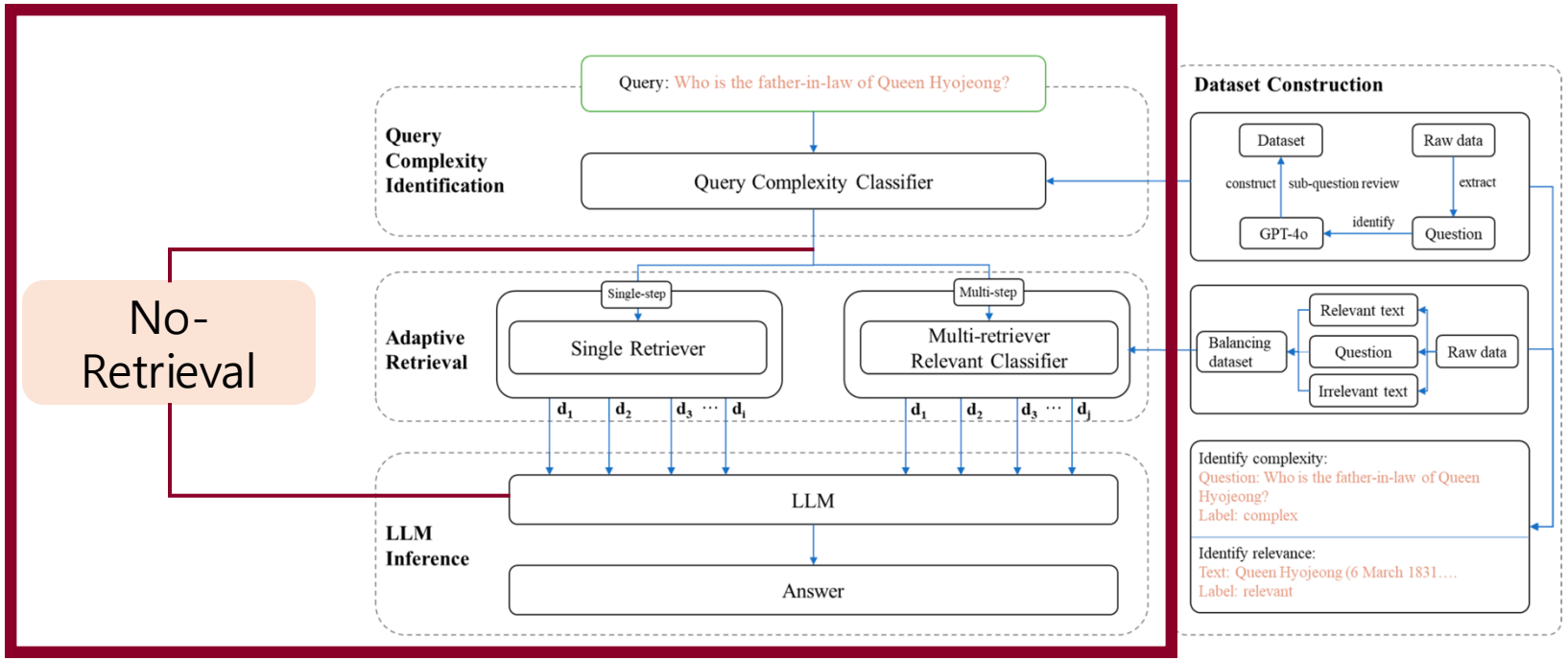
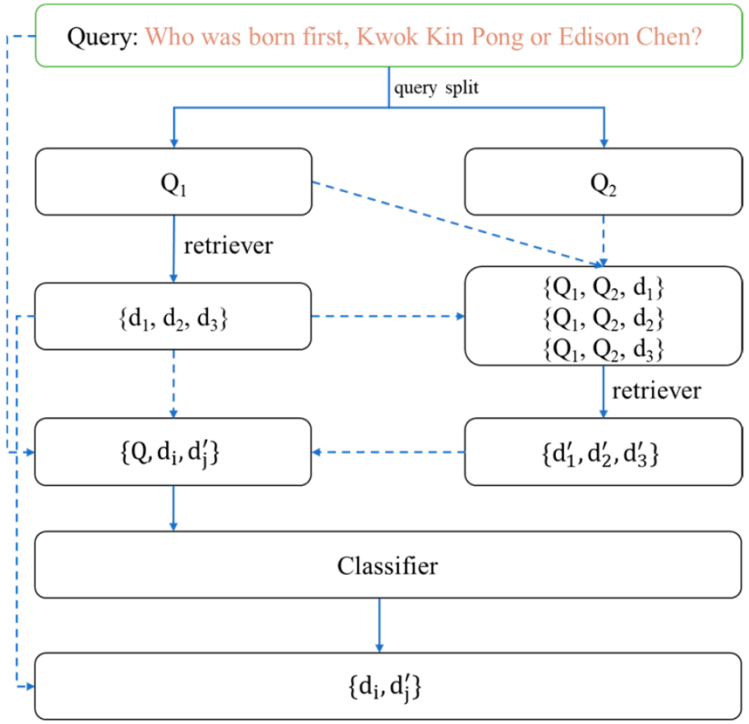


Figure 1. Workflow of the proposed LQR framework.

Query Decomposition

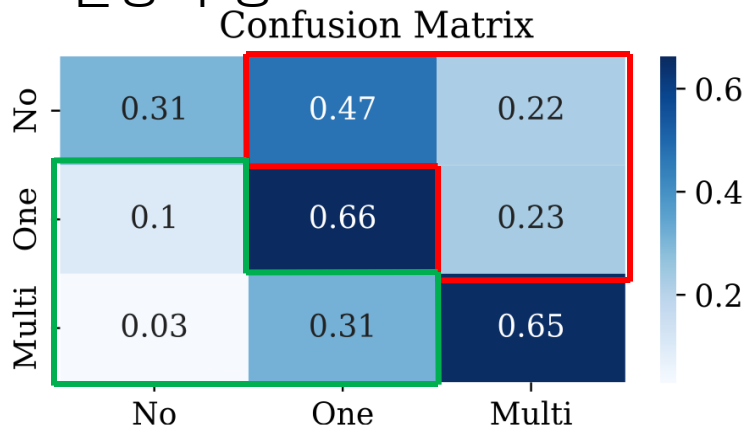


- Single-Hop 데이터셋에서는 관련 문서가 명확함에도 불구하고 Query Complexity Classifier를 활용함으로써 오히려 검색 시간이 증가하였으며, Single-Hop 데이터셋이라도 쿼리를 두개로 분리하는 경우 Efficiency 감소
- Multi-Hop 데이터셋의 경우 첫 번째 검색 결과를 기반으로 두 번째 검색에서의 문서 통합을 통해 필요한 정보들을 체계적으로 수집 및 불필요한 문서를 Relevance Classifier를 이용해 제거하여 Effectiveness, Efficiency 모두 증가

3. Future Research Direction

- 목표: 다양한 복잡도의 질의에서(특히 Multi-Hop Query) RAG를 통해 빠른 지식 검색과 높은 정확도를 유지

1. Adaptive-RAG 논문의 Confusion Matrix 정상화를 통한 Classifier Accuracy 및 QA Efficiency 향상
2. FLAN-T5가 학습에 사용한 데이터셋(SQuAD, Commonsense QA 등)을 활용한 No-Retrieval 데이터셋 구축
3. Adaptive-RAG의 Single-Hop과 Layered Query Retrieval의 Multi-Hop 방법론 병합 연구(Adaptive-RAG의 Multi-Retrieval 과정에서 각 Retrieval 단계에 Layered-Query-Retrieval의 Relevance Classifier를 도입)
4. Adaptive-RAG의 Multi-Retrieval 성능 향상을 위해 BM25 Retriever 대신 Dense Retriever 사용
5. Layered-Query-Retrieval과 같이 LLM을 이용한 Query Complexity Classifier 훈련을 위한 데이터셋의 자동 라벨링 수행



Training Strategies	QA		Classifier (Accuracy)			
	F1	Step	All	No	One	Multi
Adaptive-RAG (Ours)	**46.94**	**1084**	**54.52**	**30.52**	**66.28**	**65.45**
w/o Binary	43.43	640	60.30	62.19	65.70	39.55
w/o Silver	48.79	1464	40.00	0.00	53.98	75.91

Type	Method	SQuAD				Natural Questions				TriviaQA			
		F1	Acc	Step	Time	F1	Acc	Step	Time	F1	Acc	Step	Time
Simple	Non-Retrieval	10.50	5.00	0.00	0.11	19.00	15.60	0.00	0.13	31.80	27.00	0.00	0.13
	Single-Step Retrieval	39.30	34.00	1.00	1.00	47.30	44.60	1.00	1.00	62.40	60.20	1.00	1.00
Complex	Multi-Step Retrieval	35.60	29.60	4.52	9.03	47.80	44.20	5.04	10.18	62.40	60.20	5.28	9.22
	Adaptive Retrieval	23.10	17.60	0.50	0.55	36.00	33.00	0.50	0.56	46.90	42.60	0.50	0.56
Adaptive	Self-RAG	11.20	18.40	0.63	0.50	39.00	33.60	0.63	0.17	29.30	57.00	0.68	0.45
	Adaptive-RAG	38.30	33.00	1.37	2.02	47.30	44.60	1.00	1.00	60.70	58.20	1.23	1.54
	Our Method	**46.50**	**43.40**	3.73	4.30	**60.16**	**58.8**	3.24	3.41	**70.20**	**68.80**	3.84	4.50

Thank You

Appendix - Contribution

- 목표: 다양한 복잡도의 질의에서(특히 multi-hop query) RAG를 통해 빠른 지식 검색과 높은 정확도를 유지

1. 간단한 질의와 복잡한 질의를 모두 Adaptive 방식으로 처리하며, 효율성과 효과성 간 균형을 맞추는 Layered Query Retrieval(LQR) 프레임워크를 제공

2. 검색된 문서 중 관련 없는 문서를 필터링하여 모델 응답이 불필요한 정보의 영향을 받지 않도록 설계한 relevance classifier를 개발

3. 다수의 공개 데이터셋에 대한 실험을 통해, 본 모델이 최첨단 방법들과 비교해 Accuracy와 F1 score를 10% 이상 향상시켰음을 입증

Appendix - Query Complexity Classifier

- **Query Complexity Classifier 개요**
- 사용자의 질의가 단순한지 복잡한지 판별하기 위해, 해당 질의를 더 간단한 부분으로 분해할 수 있는지 평가
- 질의가 분해 가능한 경우 multi step query로, 분해 불가능한 경우 single step query로 분류

- **Query Complexity Classifier 구조**
- C = 0이면 단순 질의(Single retrieval), C = 1이면 복잡한 질의(multi retrieval)로 분류

$$c = Classifier(q),\ c \in \{0,1\}$$

- Linear layer와 sigmoid function으로 복잡도 평가

$$y = \sigma(Wx + b)$$

- **학습 데이터셋**
- GPT-4o와 수동 검토를 결합하여 분류기 데이터를 생성

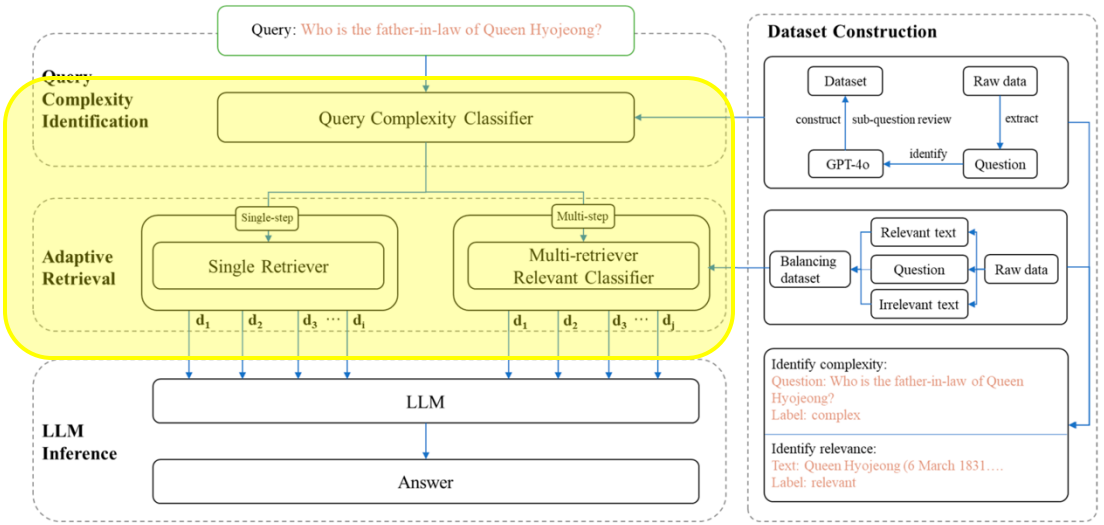


Figure 1. Workflow of the proposed LQR framework.

Appendix – Relevance Classifier

- **Single-Step Strategy**

$$rd = Retriever(q; D)$$

- efficiency: 단순한 질의를 한 번의 검색으로 처리
- accuracy: BM25 점수를 기반으로 관련 문서의 순위를 정밀하게 ranking

- **Multi-Step Strategy**

$$RD^1 = Retriever^1(q_1)$$
$$RD^2 = Retriever^2(q_1, q_2, d_i), d_i \in RD^1$$

- 검색된 문서가 질의에 기여하는지를 판별하기 위해 relevance classifier를 사용

$$rc = Classifier(q, d_j),\ d_j \in RD^2,\ rc \in \{0, 1\}$$

- rc = 0이면 문서가 질의와 관련 있음을, rc = 1이면 문서가 질의와 관련 없음을 나타낸다 -> rc = 1인 문서들은 제거

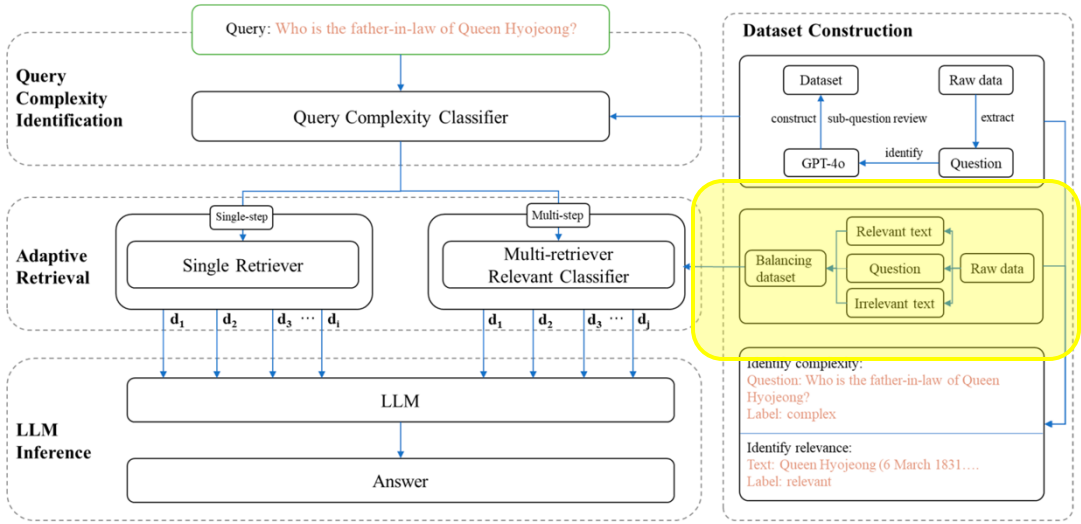
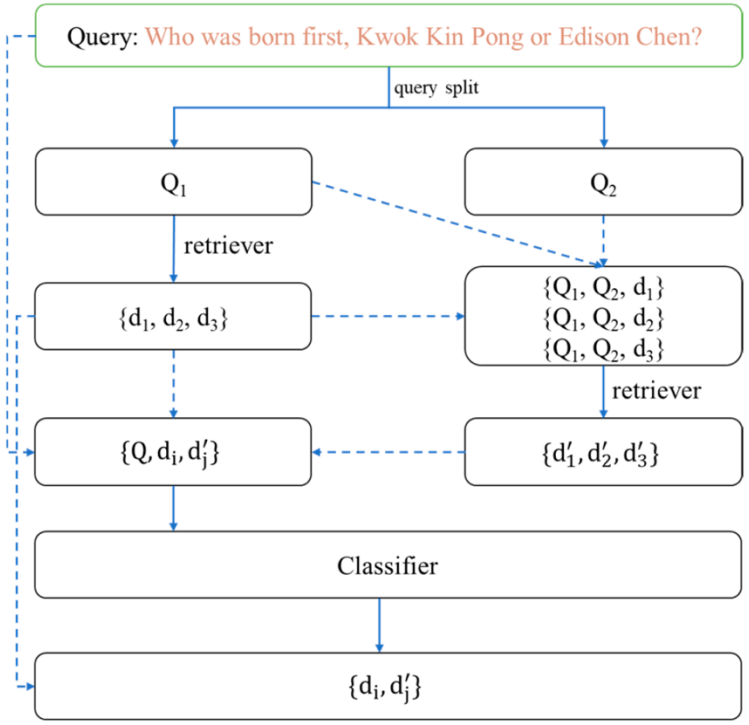


Figure 1. Workflow of the proposed LQR framework.



Appendix – Setting

- **Dataset for Evaluation**
- Single-hop dataset (SQuAD v1.1, Natural Question, TriviaQA)
- Multi-hop dataset (MuSiQue, HotpotQA, 2WikiMultiHopQA)
- Grassland sheep과 관련된 도메인 특화 데이터셋

- **Dataset for classifier**
- Query complexity identification dataset 3000개
- Relevance dataset 6000개(관련 있음 3000개, 관련 없음 3000개)

- **Evaluation Metrics**
- 효과성 평가 지표: Accuracy, F1 score
- 효율성 평가 지표: # of retrievals, time taken

- **Models for classifier**
- Query complexity identification: RoBERTa 모델
- Relevance Classifier: Longformer 모델

- **Baseline**
- Non-retrieval, single-step retrieval, adaptive-retrieval(Adaptive Retrieval, Self-RAG, **Adaptive-RAG**), multi-step retrieval

Appendix - Limitations

- **LQR 방식의 한계점**
- Multi-hop 데이터셋에서는 accuracy의 경우 모든 데이터셋에서 성능 향상을 이루었고, 효율성의 경우에도 높거나 2% 가량 떨어지므로 높은 성능 향상을 이루었다고 판단
- 반면 Single-hop 데이터셋의 경우 accuracy의 경우 Multi-hop 데이터셋보다 더 큰 성능향상을 이루었지만, efficiency의 경우 Adaptive-RAG 논문보다 상당히 떨어지며, Multi-hop 데이터셋의 수준에 불과
- **따라서 Single-hop dataset의 효율성을 향상시키기 위해 top-k sampling의 k를 1로 세팅해 초기 검색을 수행하는 방식과 첫번째 검색된 3개의 문서만으로도 하위 쿼리들이 모두 답변 가능한지 relevance classifier로 검증하는 방식을 사용**

Type	Method	SQuAD				Natural Questions				TriviaQA			
		F1	Acc	Step	Time	F1	Acc	Step	Time	F1	Acc	Step	Time
Simple	Non-Retrieval	10.50	5.00	0.00	0.11	19.00	15.60	0.00	0.13	31.80	27.00	0.00	0.13
	Single-Step Retrieval	39.30	34.00	1.00	1.00	47.30	44.60	1.00	1.00	62.40	60.20	1.00	1.00
Complex	Multi-Step Retrieval	35.60	29.60	4.52	9.03	47.80	44.20	5.04	10.18	62.40	60.20	5.28	9.22
	Adaptive Retrieval	23.10	17.60	0.50	0.55	36.00	33.00	0.50	0.56	46.90	42.60	0.50	0.56
Adaptive	Self-RAG	11.20	18.40	0.63	0.50	39.00	33.60	0.63	0.17	29.30	57.00	0.68	0.45
	Adaptive-RAG	38.30	33.00	1.37	2.02	47.30	44.60	1.00	1.00	60.70	58.20	1.23	1.54
	Our Method	**46.50**	**43.40**	3.73	4.30	**60.16**	**58.8**	3.24	3.41	**70.20**	**68.80**	3.84	4.50