2025 겨울방학 세미나
Aligning LLM to RAG framework

NLP&AI 강명훈

Natural Language Processing & Artificial Intelligence

고려대학교
KOREA UNIVERSITY

What is trustworthy RAG system?

- **LLM을 활용한 RAG system의 신뢰성(trustworthiness)을 향상하고자 할 때 고려사항**

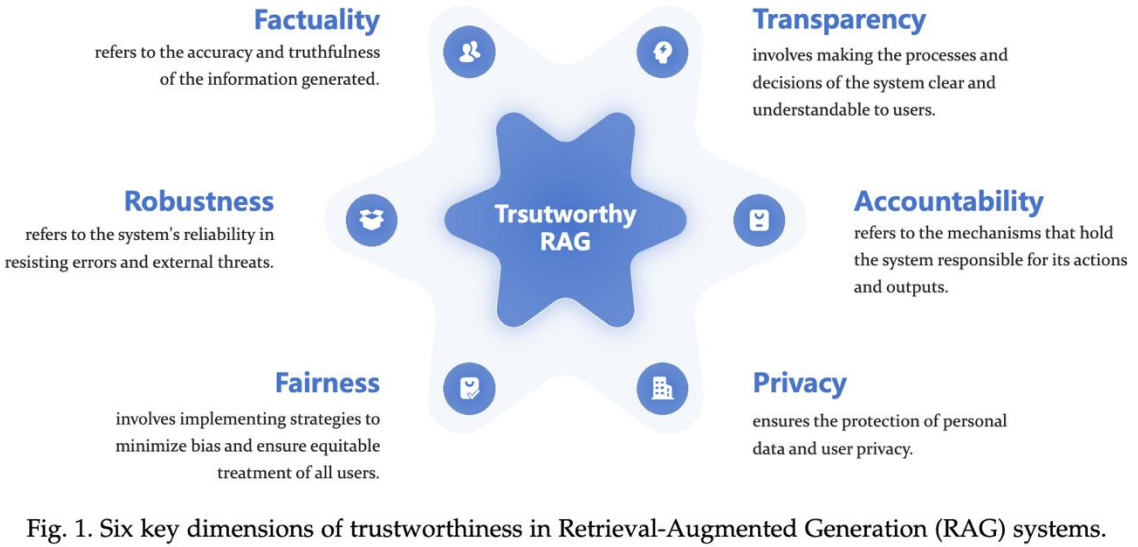


Fig. 1. Six key dimensions of trustworthiness in Retrieval-Augmented Generation (RAG) systems.

Factuality: document를 바탕으로 진실된 응답을 생성하는 능력

Robustness: adversarial attack, system error에 강건하게 대응할 수 있는 능력

Fairness: Retrieval, Generation 단계에서의 bias를 완화, 제거할 수 있는 능력

Transparency: RAG 추론 결과의 과정을 user에게 설명할 수 있는 능력

Accountability: RAG 생성 결과가 주어진 문서에서 기인하는지

Privacy: Retrieval, Generation 단계에서의 민감 정보 처리 능력

Zhou, Y., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., ... & Yu, P. S. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. arXiv preprint arXiv:2409.10102.

What is trustworthy RAG system?

- **LLM을 활용한 RAG system의 신뢰성(trustworthiness)을 향상하고자 할 때 고려사항**

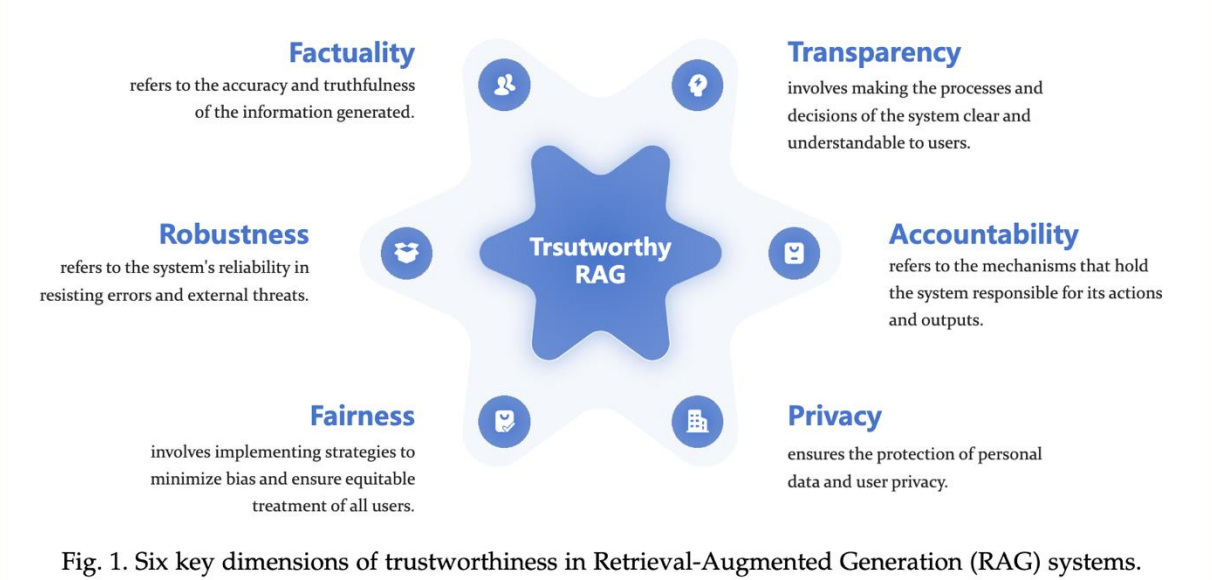


Fig. 1. Six key dimensions of trustworthiness in Retrieval-Augmented Generation (RAG) systems.

Factuality: document를 바탕으로 진실된 응답을 생성하는 능력

Robustness: adversarial attack, system error에 강건하게 대응할 수 있는 능력

Fairness: Retrieval, Generation 단계에서의 bias를 완화, 제거할 수 있는 능력

Transparency: RAG 추론 결과의 과정을 user에게 설명할 수 있는 능력

Accountability: RAG 생성 결과가 주어진 문서에서 기인하는지

Privacy: Retrieval, Generation 단계에서의 민감 정보 처리 능력

Zhou, Y., Liu, Y., Li, X., Jin, J., Qian, H., Liu, Z., ... & Yu, P. S. (2024). Trustworthiness in retrieval-augmented generation systems: A survey. arXiv preprint arXiv:2409.10102.

How can we ensure truthfulness of LM in RAG system?

- **RAG system하 LM의 trustworthiness를 향상시키는 최신의 타당한 방법을 찾아보자**

ICLR 2025 연구들을 바탕으로 RAG 상황에서의 LM의 Trustworthiness를 향상시키는 연구 탐색

Factuality 향상 연구: LM이 retrieval noise상황에서도 올바른 응답을 생성하는 denoising rationale를 생성하는 method를 제안
→ INSTRUCTRAG: INSTRUCTING RETRIEVAL AUGMENTED GENERATION VIA SELF-SYNTHESIZED RATIONALES

Accountability 향상 연구: LM이 주어진 document만 활용하여 응답할 수 있는 능력을 정확하게 측정 및 향상하는 method를 제안
→ MEASURING AND ENHANCING TRUSTWORTHINESS OF LLMS IN RAG THROUGH GROUNDED ATTRIBUTIONS AND LEARNING TO REFUSE

Truthfulness of LM in RAG

INSTRUCTRAG: INSTRUCTING RETRIEVAL AUGMENTED GENERATION VIA SELF-SYNTHESIZED RATIONALES

Zhepei Wei **Wei-Lin Chen** **Yu Meng**
Department of Computer Science
University of Virginia
{zhepei.wei,wlchen,yumeng5}@virginia.edu

ICLR 2025 poster

score: 8 / 8 / 5 / 6 / 8

Why should we refer this paper?

- **Retrieval error and Factuality**

RAG system에서 external knowledge는 retriever로 획득. 이 때 retriever로 external knowledge를 얻는 과정은 다음의 위험성을 내포함

- Retriever의 performance에 따라서 질의와 관련이 없는 document가 추출될 가능성
- Open-domain 상황에서 Retriever의 검색 pool이 질의와 관련이 없을 가능성

← Imperfect retrieval 결과에 질의와 관련이 없는 document가 존재할 가능성 상시 존재

RAG system에서 external knowledge를 바탕으로 답변을 생성하는 LM의 retrieval error에 강건한 능력 필요

→ LLM의 reasoning 능력을 바탕으로 검색 external knowledge가 주어질 때

1) 답변 생성에 필요한 document를 찾아내는 과정인 설명력 있는 rationale를 생성하고

2) 해당 rationale를 바탕으로 demonstration, training set을 구성하여 retrieval error에 강건한 LM을 기획하자

추가적인 Supervision 없이 LLM의 rationale 생성 결과를 바탕으로 retrieval error에 강건한 trainable, training-free 전략을 모두 구축하는 InstructRAG를 제안

InstructRAG

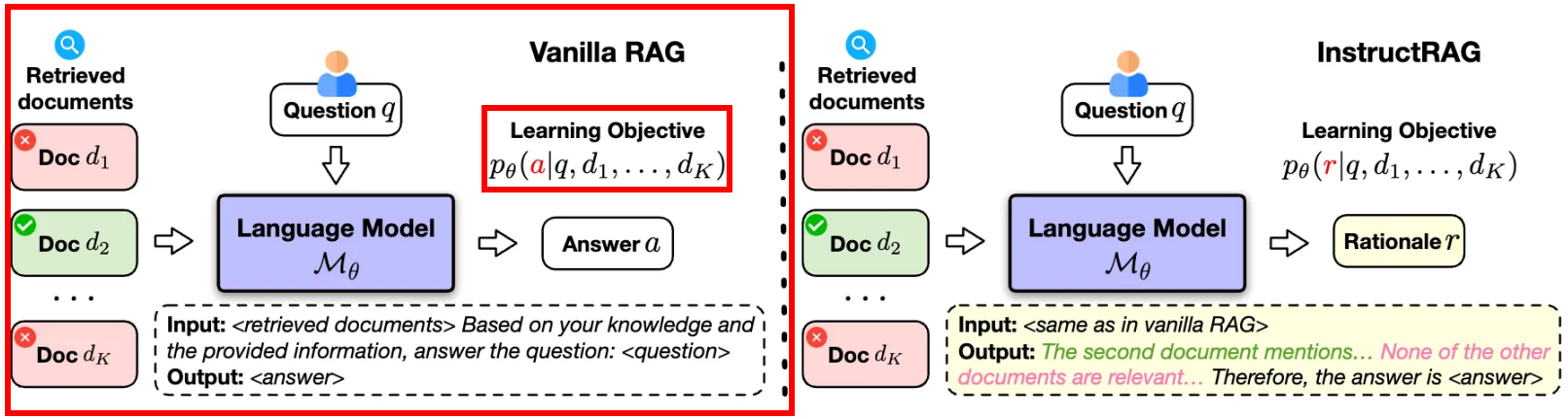


Figure 1: Comparison between vanilla RAG and our INSTRUCTRAG. In vanilla RAG, the model is tasked to directly predict answers given user queries and potentially noisy retrieved documents, without explicit denoising processes or explanations for how the answer is derived. In contrast, our proposed INSTRUCTRAG generates rationales that explicitly denoise the retrieved documents and justify the predicted answers, enhancing both the generation accuracy and trustworthiness.

Retrieval error가 상주하는 RAG system하에서 answer generation을 구성하는 Vanila RAG system의 vulnerability 지적

- 기존의 방법들은 noise가 존재하는 상황에서 direct answer generation을 하도록 implicit하게 학습
- 이러한 implicit 학습 방식은 LM의 output의 추론 근거, 오류 원인을 파악하기 어려움

최근 이러한 Vanila RAG system의 answer generation 방식의 오류를 해결하고자 answer 생성 과정에 denoising을 하는 방법이 기획되고 있음

- Denoising은 answer 생성 과정에 LM이 주어진 document set중 필요한 document와 그렇지 않은 것을 구분하는 rationale를 생성하는 것

← 이러한 Denoising을 직접 학습하도록 dataset을 구축할 수 있으나 human annotation cost가 매우 높은 실정

InstructRAG

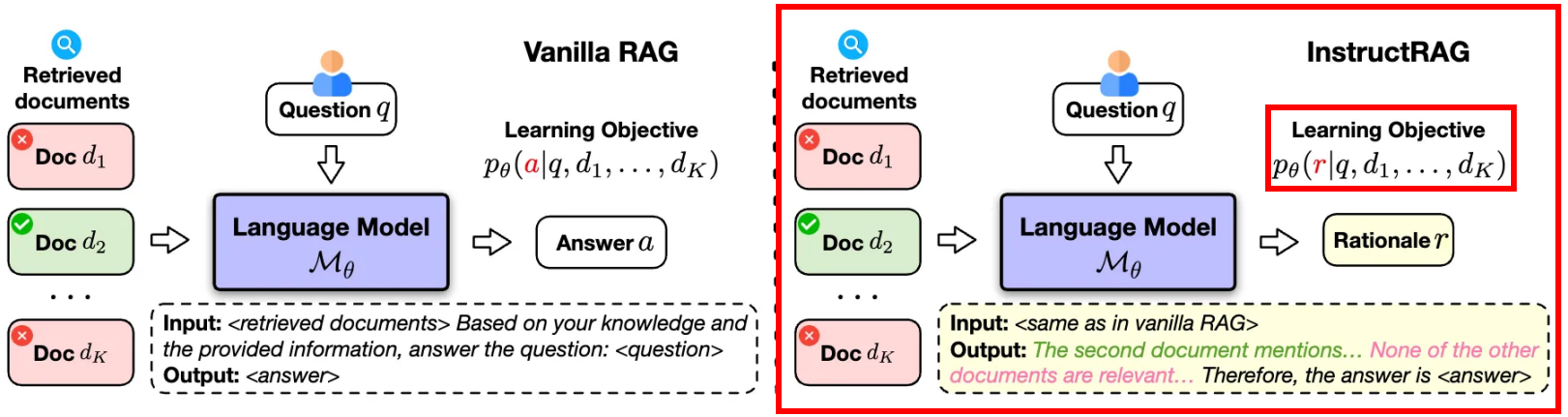


Figure 1: Comparison between vanilla RAG and our INSTRUCTRAG. In vanilla RAG, the model is tasked to directly predict answers given user queries and potentially noisy retrieved documents, without explicit denoising processes or explanations for how the answer is derived. In contrast, our proposed INSTRUCTRAG generates rationales that explicitly denoise the retrieved documents and justify the predicted answers, enhancing both the generation accuracy and trustworthiness.

RAG system에 상주하는 retrieval error를 고려한 denoising을 제안하는 InstructRAG를 제안

- 효과적인 denoising을 위해 Explicit하게 denoising rationale를 생성하는 방법 제안
  - 생성된 결과물의 정답여부 (verifiability), 신뢰성 (trustworthiness)를 보장할 수 있음
- 효율적인 denoising rationale 학습을 위한 self-synthesized dataset 제안
  - 학습 대상이 되는 LM의 rationale를 바탕으로 학습을 진행하여 human annotation cost를 줄일 수 있음

InstructRAG – Method

- **Step 1: Rationale Generation**

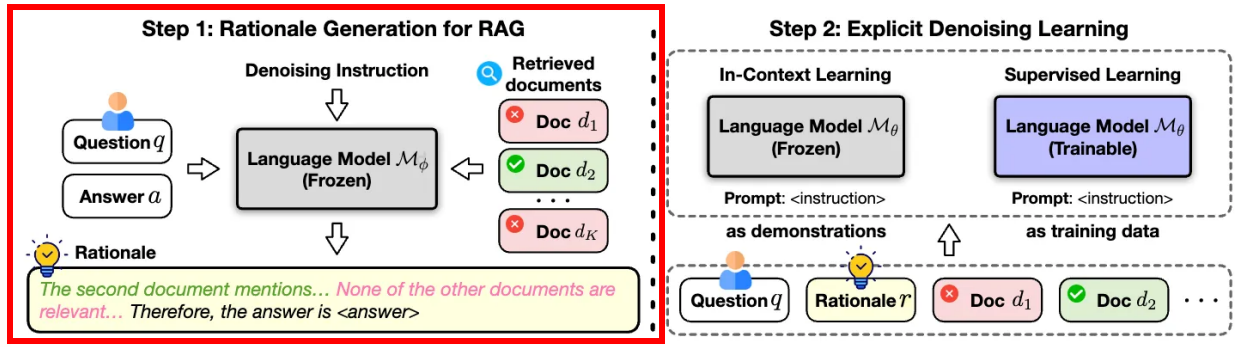


Figure 2: An overview of INSTRUCTRAG. In step one, given the question q, retrieved documents $\{d_1, \cdots, d_K\}$ and ground-truth answer a from the training set, we prompt an instruction-tuned LM (*i.e.*, rationale generator $\mathcal{M}_\phi$) to generate rationale r that explains how the answer can be derived from the potentially noisy input. In step two, we utilize the synthesized rationales from the first step to guide the LM (*i.e.*, rationale learner $\mathcal{M}_\theta$) to explicitly learn denoising of the retrieved documents, either through in-context learning or supervised learning. By default, we use the same model for both $\mathcal{M}_\phi$ and $\mathcal{M}_\theta$, but they can be instantiated with different models as well (see ablation study § 3.3).

Explicit하게 rationale를 generation하는 기반 작업으로 rationale r 생성

- RAG 상황을 요구하는 원본 데이터셋의 output을 증강하는 방법으로 접근

for each $\langle q, a \rangle \in \mathcal{T}$ **do**
  Retrieve $D = \{d_1, \cdots, d_K\} \leftarrow \mathcal{R}(q)$
  Synthesize denoising rationale $r \leftarrow \mathcal{M}_\phi(q, a, D)$
Augment training data $\mathcal{T} \rightarrow \mathcal{T}^+ = \{\langle q, r \rangle\}$

- substring match로 r 과 a 의 일치율이 98%이므로 두 text간의 consistency 보장

Notation

$\mathcal{M}_\phi$ rationale를 생성하는 LM

$\mathcal{R}$ off-the-shelf retriever

$\mathcal{T} = \{\langle q, a \rangle\}$ 질의 q와 GT answer a로 구성된 원본 데이터셋

$\mathcal{T}^+ = \{\langle q, r \rangle\}$ 질의 q와 denoising rationale, answer가 포함된 증강된 데이터셋

InstructRAG – Method

- **Step 2: Explicit Denoising Learning**

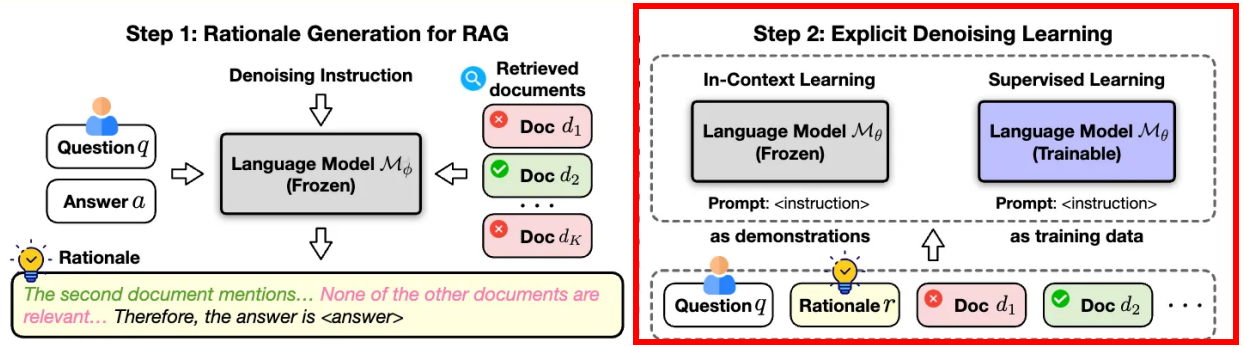


Figure 2: An overview of INSTRUCTRAG. In step one, given the question q, retrieved documents $\{d_1, \cdots, d_K\}$ and ground-truth answer a from the training set, we prompt an instruction-tuned LM (*i.e.*, rationale generator $\mathcal{M}_\phi$) to generate rationale r that explains how the answer can be derived from the potentially noisy input. In step two, we utilize the synthesized rationales from the first step to guide the LM (*i.e.*, rationale learner $\mathcal{M}_\theta$) to explicitly learn denoising of the retrieved documents, either through in-context learning or supervised learning. By default, we use the same model for both $\mathcal{M}_\phi$ and $\mathcal{M}_\theta$, but they can be instantiated with different models as well (see ablation study § 3.3).

Notation

$\mathcal{M}_\theta$ 질의와 문서 set이 입력으로 주어질 때 rationale를 생성하도록 학습된 LM

LM이 noise가 포함된 document set이 주어질 때에도 denoising rationale를 생성하는 explicit denoising learning을 실시

if `LearningMode == In-Context Learning` **then** ▷ INSTRUCTRAG-ICL
  Sample ICL examples $\mathcal{E} = \{\langle q, r\rangle\} \subseteq \mathcal{T}^+$
  $r \leftarrow \mathcal{M}_\theta(r|q, \mathcal{R}(q), \mathcal{E})$ given inference query q ▷ Detailed in Table 10
else if `LearningMode == Fine-Tuning` **then** ▷ INSTRUCTRAG-FT
  Fine-tune $\mathcal{M}_\theta$ on $\mathcal{T}^+$ with retrieved documents $\{\langle q, r, D\rangle\}$
  $r \leftarrow \mathcal{M}_\theta(r|q, \mathcal{R}(q))$ given inference query q ▷ Detailed in Table 11
return r

InstructRAG-ICL

- (question, answer, rationale)를 n개의 shot으로 주고 question, document set이 주어질 때 Denoising rationale을 생성하도록 In-context Learning

InstructRAG-FT $\mathcal{T}^+ = \{\langle q, r\rangle\}$

- step1 에서 생성한 를 학습 데이터로 이용하여 explicit하게 LLM이 denoising rationale를 생성하도록 학습
- 새로운 loss 설계를 하지 않고 일반적으로 사용되는 $\max_\theta \mathbb{E}_{(q,r)\sim\mathcal{T}^+} \log p_\theta(r|q, D)$.

InstructRAG – Experimental Setups

- **Datasets**

Table 2: Dataset statistics and retrieval setting.

Dataset	Train	Test	Retriever	Top-K	Recall@K
PopQA	12,868	1,399	Contriever	5	68.7
TriviaQA	78,785	11,313	Contriever	5	73.5
Natural Questions	79,168	3,610	DPR	5	68.8
ASQA	4,353	948	GTR	5	82.2
2WikiMultiHopQA	167,454	12,576	BM25	10	40.7

Knowledge-intensive benchmark에 관하여 InstructRAG의 우수성 증명

- **PopQA**: wikipedia에 존재하는 entity를 바탕으로 QA pair를 구성하여 popular, unpopular 지식에 따른 LM의 응답 양상 측정하는 open-domain QA dataset
- **Trivia QA**: 전체 evidence내 일부분인 trivia question에 대하여 answering을 요구하는 MRC dataset
- **Natural Questions**: 사용자들의 google search query를 바탕으로 구성한 QA dataset
- **ASQA**: 관점에 따라 여러 답변이 존재하는 ambiguous question에 관하여 답변을 생성해야 하는 long-form QA dataset
- **2WikiMultiHopQA**: Wikipedia, Wikidata를 바탕으로 구성된 Multi-hop QA dataset. 답변 생성의 근거를 triplet 형태로 제공

InstructRAG – Experimental Setups

- **Datasets**

Table 2: Dataset statistics and retrieval setting.

Dataset	Train	Test	Retriever	Top-K	Recall@K
PopQA	12,868	1,399	Contriever	5	68.7
TriviaQA	78,785	11,313	Contriever	5	73.5
Natural Questions	79,168	3,610	DPR	5	68.8
ASQA	4,353	948	GTR	5	82.2
2WikiMultiHopQA	167,454	12,576	BM25	10	40.7

각 dataset마다 retriever에서 얻은 검색 document set을
LM에게 입력으로 주는 방식으로 실험 진행

- dataset마다 Recall@k가 100에 가깝지 않은 점은 논문에서 제기하고 있는 imperfect retrieval result 상황을 말하는 것

InstructRAG – Experimental Setups

- **Metric & baselines**

Metric

- **Accuracy:** PopQA, Trivia QA, Natural Questions, 2WikiMultiHopQA
- **Exact Match (EM)**: ASQA
- **citation precision (pre), recall (rec)**: ASQA
  - NLI 모델을 활용하여 모델이 생성한 답변 중 statement s_i, Citation set C_i 간의 entailment를 바탕으로 precision, recall 측정
  - recall: 전체 C_i가 statement s_i 를 entail 하는지 → s_i 가 document set에 근거해서 생성되었는지 확인
  - precision: 개별 $c_{i,j}$ 가 statement s_i 를 entail 하는지 → 필요한 citation만 수행하였는지 확인

Baseline

제안하는 방식이 training-free, trainable setting이기 때문에 baselines도 이에 맞춰 선정

1. training-free setting
- Vanilla zero-shot prompting
  - non-retrieval zero-shot baseline
- Few-shot demonstration with instruction
  - non-retrieval few-shot baseline
- RALM : in-context retrieval-augmented language modeling
  - retrieval document + query 입력으로 answer generation 진행

2. trainable setting
- vanilla supervised fine-tuning (SFT)
- RetRobust
- Self-RAG

InstructRAG – Experimental Setups

- **baselines details**

RetRobut

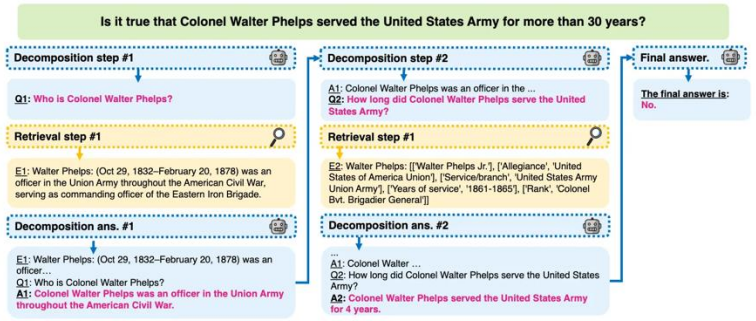


Figure 3: Interleaving decomposition and retrieval in Self-Ask format (Press et al., 2023). The model generates intermediate questions and answers until generating the final answer (model generations are shown in pink). Retrieved evidence for intermediate questions is prepended at each step.

- RAG상황에 학습되지 않은 LLM이 retrieval 결과를 활용할 때 robust하도록 학습 방법을 제안. RALM처럼 LLM에게 $\{D, Q\}$입력이 주어질때 A를 생성하는 상황을 학습시킴
- single-hop setting: 기존 dataset에 off-the-shelf retrieval을 이용해 top-1 retrieval context, low-ranked retrieval context를 합쳐 r_q, q 를 주고 답변 a를 생성하도록 지시
- multi-hop setting: multi-hop QA를 위해 한 개의 question당 2~4개의 decomposition question을 생성한 뒤 각 decomposed question마다 retrieval을 실시.

SelfRAG

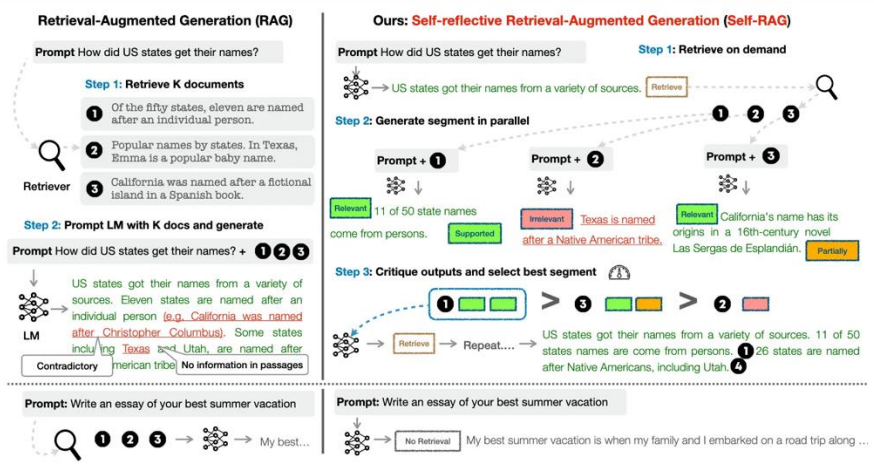


Figure 1: Overview of SELF-RAG. SELF-RAG learns to retrieve, critique, and generate text passages to enhance overall generation quality, factuality, and verifiability.

LLM에게 query가 주어질 때 retrieval이 필요한지, 그리고 필요하다면 주어진 document에 따라 생성하는 결과가 relevant한지, supportive한지, useful한지를 평가하면서 답변을 생성하는 Self-RAG generator를 생성

Inference의 경우 아래와 같은 방식으로 실시

- retrieval이 필요한 경우 off-the-shelf retriever을 이용하여 top-k retrieval 실시
- 각 document마다 parallel하게 beam-search generation을 실시
  - beam-search generation을 할 때 top-B의 segment를 생성
  - 이 때 top-B를 선정하는 기준은 IsRel, IsSup, IsUse token 생성확률을 기준으로 선정

InstructRAG – Result

- **Main result**

Table 3: Overall results of INSTRUCTRAG and baselines on five knowledge-intensive benchmarks in training-free and trainable RAG settings. We re-implement baselines and report their performance as the higher one between the original scores and our reproduced results. * indicates the results copied from Asai et al. (2023b) for reference. "–" indicates the results are not reported in the original paper or not applicable (*e.g.*, some methods cannot produce citations). The best performance is highlighted in ***bold***.

Method	PopQA (acc)	TriviaQA (acc)	NQ (acc)	MultiHopQA (acc)	ASQA (em)	ASQA (pre)	ASQA (rec)
Baselines w/o Retrieval							
Vanilla Zero-shot Prompting							
ChatGPT*	29.3	74.3	–	–	35.3	–	–
Llama-3-Instruct$_{8B}$	22.8	69.4	46.6	45.6	30.6	–	–
Llama-3-Instruct$_{70B}$	28.9	80.6	57.9	57.5	39.1	–	–
RAG w/o Training							
In-Context RALM (Ram et al., 2023)							
ChatGPT*	50.8	65.7	–	–	40.7	65.1	76.6
Llama-3-Instruct$_{8B}$	62.3	71.4	56.8	43.4	40.0	62.1	66.4
Llama-3-Instruct$_{70B}$	63.8	76.3	60.2	51.2	43.1	62.9	67.6
Few-Shot Demo. w/ Instruction							
Llama-3-Instruct$_{8B}$	63.1	74.2	60.1	45.3	42.6	55.0	64.4
Llama-3-Instruct$_{70B}$	63.9	79.1	62.9	53.9	45.4	49.3	57.1
INSTRUCTRAG-ICL							
Llama-3-Instruct$_{8B}$	64.2	76.8	62.1	50.4	44.7	**70.9**	**74.1**
Llama-3-Instruct$_{70B}$	**65.5**	**81.2**	**66.5**	**57.3**	**47.8**	69.1	71.2
RAG w/ Training							
Vanilla Supervised Fine-tuning							
Llama-3-Instruct$_{8B}$	61.0	73.9	56.6	56.1	43.8	–	–
Self-RAG (Asai et al., 2023b)							
Llama-2$_{7B}$	55.8	68.9	42.4	35.9	30.0	66.9	67.8
Llama-2$_{13B}$	56.3	70.4	46.4	36.0	31.4	**70.3**	**71.3**
Llama-3-Instruct$_{8B}$	55.8	71.4	42.8	32.9	36.9	69.7	69.7
RetRobust (Yoran et al., 2024)							
Llama-2$_{13B}$	–	–	39.6	51.5	–	–	–
Llama-3-Instruct$_{8B}$	56.5	71.5	54.2	54.7	40.5	–	–
INSTRUCTRAG-FT							
Llama-3-Instruct$_{8B}$	**66.2**	**78.5**	**65.7**	**57.2**	**47.6**	65.7	70.5

between the original scores and our reproduced results. As our method adopts instruction-tuned Llama-3 as the backbone model, we also train RetRobust and Self-RAG with Llama-3-Instruct$_{8B}$ and optimize their performance through extensive hyper-parameters search. More details on implementation, including training, inference, and prompt design are available in Appendix B and Appendix D. We also present some case studies in Appendix C.

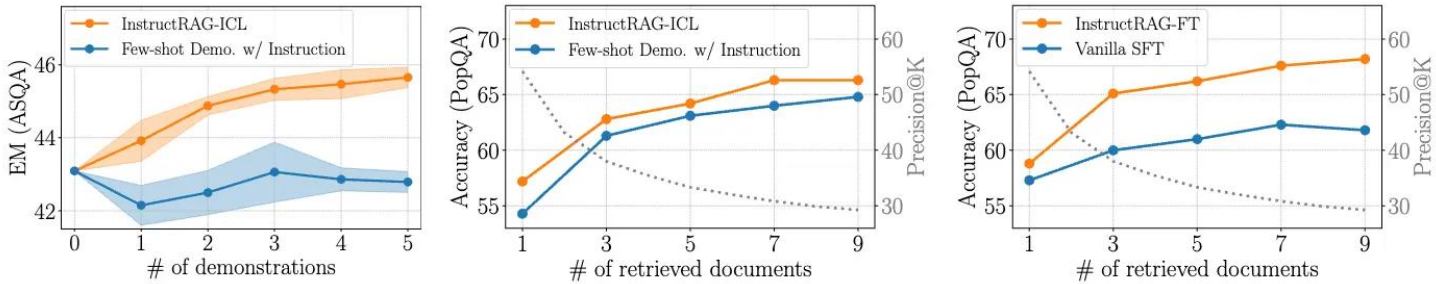
- Without retrieval: TriviaQA에서 높은 성능이 달성됨
  → parametric knowledge로 충분히 달성할 수 있는 task, data contamination 이슈 제기

- RAG without training: 전반적으로 Non-retrieval baseline 대비 높은 성능
  + InstructRAG가 최고성능 달성

- RAG with training: ASQA citation에서 self-rag 대비 떨어지는 성능 제외 모두 최고 성능 달성
  - 이는 제안하는 방법이 denoising에 초점을 맞추기 있기 때문에 citation 관련 성능이 떨어지는 것으로 기인. 그러나 여전히 comparable한 성능 달성
  - 그러나 Self-RAG는 다른 데이터셋에서 타 training baseline 특히 SFT 대비 우수한 성능을 보이지 못함
  - NQ, PopQA, TriviaQA의 경우 일반 baseline및 w/o training과 대비하여 낮은 성능을 보이는 경우도 존재 → 낮은 generalizability
  - RetRobust는 MultiHopQA에서 comparable한 성능달성.
  이는 해당 method가 multi-hop QA를 고려한 training 방법에서 기인한 것으로 추정

InstructRAG – Result

- **Robustness of InstructRAG with respect to noise ratios**



(a) Training-free RAG setting. (b) Training-free RAG setting. (c) Trainable RAG setting.

Figure 3: Impact of different number of demonstrations and retrieved documents. (a) Demonstration sensitivity study of INSTRUCTRAG-ICL. (b) Noise robustness study of INSTRUCTRAG-ICL. (c) Noise robustness study of INSTRUCTRAG-FT.

- ASQA, PoP-QA 데이터셋에 대해서 Top-k document를 늘림에 따라 성능 변화를 파악함

- 제안한 InstructRAG가 baseline인 few-shot 대비 demonstration 증가에 따른 일관적인 성능 향상을 보임

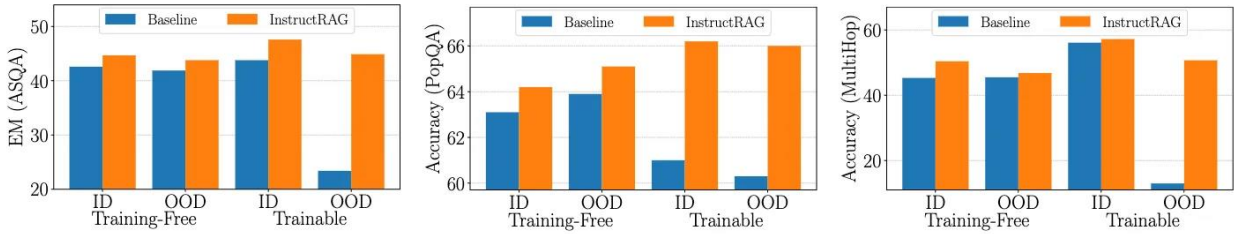
→ Direct answer generation을 요구하는 baseline 대비 rationale gen demonstration 효과성 입증

- 제안한 InstructRAG가 baseline인 few-shot 및 SFT 대비 noise 개수에도 강건한 모습을 보임 (b), (c)

→ trainable, training-free 모든 상황에서 noise ratios 증가에도 robust한 성능을 보임

InstructRAG – Result

- **Generalizability of InstructRAG with respect to unseen tasks**



(a) Short-form to long-form QA. (b) Long-form to short-form QA. (c) Single-hop to multi-hop QA.

Figure 4: Generalizing INSTRUCTRAG from source domain task to target domain task, where ID and OOD denote in-domain and out-of-domain settings. (a) PopQA (short-form QA task) as source domain and ASQA (long-form QA task) as target domain. (b) ASQA as source domain and PopQA as target domain. (c) PopQA (single-hop QA task) as source domain and 2WikiMultiHopQA (multi-hop QA task) as target domain. We adopt *few-shot demonstration with instruction* and *vanilla supervised fine-tuning* as the training-free and trainable baselines.

- InstructRAG와 baseline을 in-domain data(ID)에 학습을 수행하거나 ICL shot에 in-domain data를 주고 Out-of-domain data(ood)에 관하여 성능을 측정
  - Long-form QA: ASQA task, Short-form QA, single-hop QA: PopQA, Multi-hop QA: 2WIkiMultiHopQA
  - training-free의 경우 in-domain demonstration을 가지고 ood에 적용했을 때 파악

- 제안한 InstructRAG가 baseline인 few-shot 및 SFT 대비 보다 unseen task generalization능력을 보임

Truthfulness of LM in RAG

MEASURING AND ENHANCING TRUSTWORTHINESS OF LLMS IN RAG THROUGH GROUNDED ATTRIBUTIONS AND LEARNING TO REFUSE

Maojia Song*, **Shang Hong Sim***, **Rishabh Bhardwaj**
Singapore University of Technology and Design
{maojia_song, shanghong_sim, rishabh_bhardwaj}@mymail.sutd.edu.sg

Hai Leong Chieu
DSO National Laboratories
chaileon@dso.org.sg

Navonil Majumder, Soujanya Poria
Singapore University of Technology and Design
{navonil_majumder, sporia}@sutd.edu.sg

ICLR 2025 oral

score: 8 / 8 / 8 / 8

Why should we refer this paper?

- **Grounding error and Accountability**

  RAG system은 검색된 external knowledge에 존재하는 정보를 grounding하여 질의에 관련한 답변을 생성하는 시스템임
  이상적인 RAG system의 LM은 오직 주어진 document set에 근거해서 답변을 생성 해야함. 이 때 다음의 사항들이 모두 고려되어야 함

  - 주어진 document set이 질의 답변에 불충분할 경우 답변 생성을 진행하지 않아야 함 (**refusal**)
  - 주어진 document set을 통해 생성된 답변이 document의 어떤 부분에 기인해 생성되었는지 파악할 수 있어야 함 (**citation**)
  - document의 특정 부분에 기인해 생성했더라도 해당 내용이 정확한지 파악되어야 함 (**correctness**)

  또한 LM이 parametric-knowledge를 사용하여 external knowledge를 grounding 하지 않고 옳은 답변을 수행하는 경우에 관한 penalty 부여 필요

  RAG system 하 LM의 grounding 능력을 다방면으로 평가하는 metric 필요
  → RAG system에서 Retriever, Parametric knowledge가 관여하는 부분을 제외하여 평가를 진행하자

  RAG system에서 LM의 document set grounding 능력을 refusal, citation, correctness 모두를 고려한 TRUST-SCORE metric 제안 +
  LM의 grounding 능력 및 citation능력을 향상할 수 있는 TRUST-ALIGN dataset을 제안

TRUST-SCORE

- **Motivation**
  - 본 연구는 LLM이 주어진 일련의 document 정보만을 활용하여 질의에 관한 답변을 수행하는 LLM의 document grounding 능력을 탐색하고자 함

  - 이때 RAG system 하에서 LLM의 grounding 능력을 파악하고자 할 때 LLM이 생성한 output을 바탕으로 주어진 document 정보에 근거해서 생성했는지 여부를 파악하고자 함

  - LLM의 output을 바탕으로 groundedness를 파악하고자 할 때 LM 앞 단의 retriever의 영향력과 LM의 parametric knowledge의 영향력을 고려하여 판단해야함
    - Retriever의 영향력 제외: 정답 생성에 관여하는 document를 적절히 **citation하는지 평가**
      ← document set 중 관련 있는 document를 적절히 사용했는지에 관한 평가
    - Parametric knowledge 영향력 제외: 주어진 document set으로 응답을 할 수 없는 경우 **refusal 능력 평가**
      ← LLM의 parametric knowledge를 사용하는 것이 아닌 주어진 external knowledge를 얼마나 잘 활용하는지에 관한 문제

TRUST-SCORE

- **Motivation**
  - 본 연구는 RAG 상황에서 LLM의 trustworthiness를 다방면에서 고려하는 metric인 TRUST-SCORE를 제안

  - TRUST-SCORE는 3가지 측면에서 RAG system 하 LM의 Groundedness를 평가
    - **Refusal**: The ability to discern which questions can be answered or refused based on the provided documents
    - Calibrated **Correctness**: Claim recall scores for the answerable questions
    - **Citation**
      - Recall: The extent to which generated claims are supported by the corresponding citations
      - Precision: The relevance of the citations to the statements

  → TRUST-SCORE is designed to specifically measure the LLM's performance within a RAG setup,
  isolating it from the influence of retrieval quality.

TRUST-SCORE

- **Problem Setting**
  - 본 연구는 LLM이 주어진 document set으로만 grounding해서 답변을 해야 하는 상황만을 다룸
    → 따라서 LLM의 parametric knowledge로 document set에 없는 내용을 답변하는 것은 hallucination에 해당함
  - 본 연구는 LLM의 trustworthiness, 신뢰성을 groundedness로 치환하여 접근

  * Groundedness 측정에서 고려해야 할 3가지 요소

  Answerability: question이 주어진 document set에 entail 되는 경우

  Refusal: question이 주어진 document set으로 답변이 불가능한 경우

  Hallucination: LLM의 답변이 주어진 document set에 기인하지 않은 경우
  - Inaccurate answer → $\mathrm{EM}^{\mathrm{F1}}_{\mathrm{AC}}$
  - Over-Responsiveness: refusal로 답해야하는 상황에서 **답변을 생성**하는 경우 $\mathrm{F1}_{\mathrm{RG}}, \mathrm{F1}_{\mathrm{ref}}$
  - Excessive Refusal: 답변을 할 수 있는 상황에서 **refusal로 답변을 생성**하는 경우 $\mathrm{F1}_{\mathrm{RG}}, \mathrm{F1}_{\mathrm{ans}}$
  - Overaction: 모델이 답변을 생성하나 **불필요한 citation**을 생성하는 경우 $\mathrm{F1}_{\mathrm{CG}}, \mathrm{CP}$
  - Improper Citation: 모델이 답변 생성에 참조한 **citation이 답변 생성에 필요 없는 경우** $\mathrm{F1}_{\mathrm{CG}}, \mathrm{CR}$

  → 즉 Grounding 능력이 높은 LLM은 주어진 document set으로 답변을 할 수 있는 상황에서만 답변을 수행하고 그 중 관련 있는 document만을 응답 생성에 사용해야 함

TRUST-SCORE

- **Problem Setting**

Notation

$A_G = \{a_{g1}, \ldots, a_{gn}\}$ gold answer를 **claim 단위로 분해**한 것

$A_D = \{a_{d1}, \ldots, a_{dn}\}$ gold claim중 주어진 **document set에서 기인할 수 있는 claim**

$A_R = \{a_{r1}, \ldots, a_{rn}\}$ 모델이 생성한 답변 **response를 claim 단위로 분해**한 것

기존의 groundedness metric 연구는 다음의 식으로 구성됨 $\mathrm{EM}_{\mathrm{r}} = \frac{|A_D \cap A_G|}{|A_G|} + \frac{|A_R \cap A_G|}{|A_G|}$

이는 LM의 parametric-knowledge가 개입하는 부분 $\frac{|A_R \cap A_G|}{|A_G|}$ 을 고려하지 못함

본 논문은 parametric-knowledge 개입이 제거된 상황의 식 $\mathrm{EM}_{\mathrm{r}} = \frac{|A_D \cap A_G|}{|A_G|}$ 으로 groundedness upper bound를 설정하고 metric을 구성

TRUST-SCORE

- **Overview**

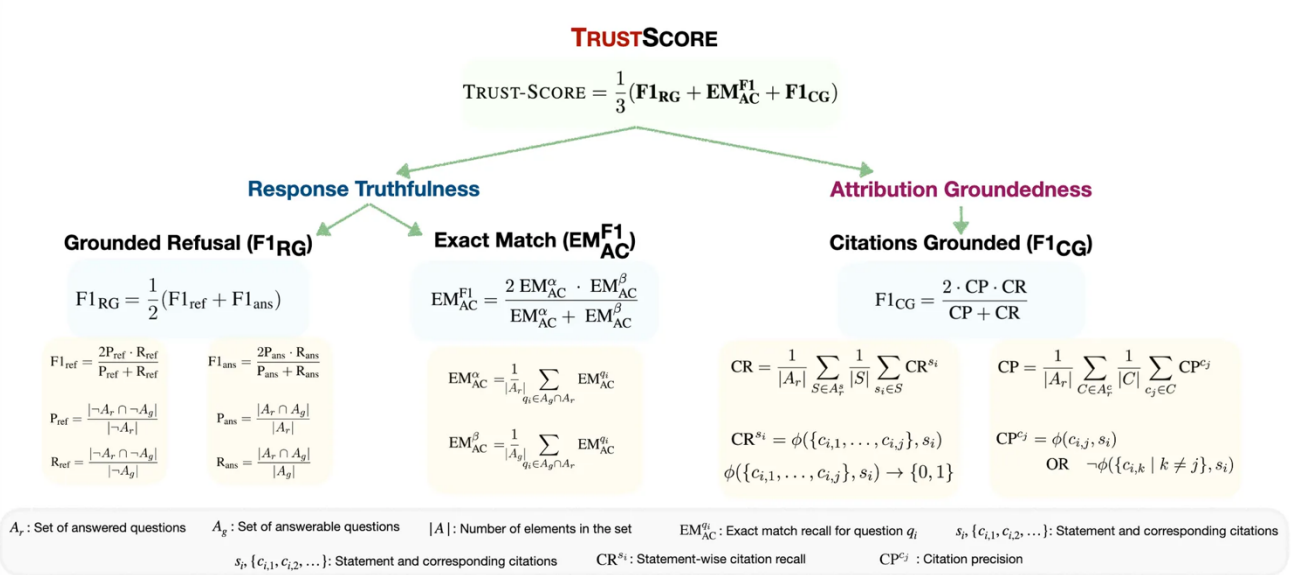


Figure 1: TRUST-SCORE calculation shown as a computational graph.

TRUST-SCORE는 LLM을 활용한 RAG system의 Truthfulness, Groundness를 동시에 고려하여 하나의 점수로 산출함

- Response Truthfulness: 생성된 응답이 진실한지
- Attribution Groundedness: 생성된 응답에 포함된 citation이 제대로 생성되었는지

$$\text{Trust-Score} = \frac{1}{3}(\text{F1}_{\text{RG}} + \text{EM}^{\text{F1}}_{\text{AC}} + \text{F1}_{\text{CG}})$$

TRUST-SCORE

- **Response Truthfulness: Exact Match Recall**

  모델이 생성한 부분 중에 document에 기인하여 생성했으면서 그 부분이 사실에 부합한지에 관한 문제 (Correctness)

  - 모델이 생성한 response 중 document D에 기인하면서 correct한지 여부를 recall 관점으로 측정
  - 계산의 기준이 되는 전체 case는 응답할 수 있는 question set A_g, 모델이 응답한 question set A_r

  * Exact Match Recall $\mathrm{EM}_{\mathrm{AC}}^{q_i} = \frac{|A_G \cap A_D \cap A_R|}{|A_G \cap A_D|}$

  - sample 별 EM recall

  - 전체 dataset 차원에서 EM recall은 아래와 같이 구성가능

    $\mathrm{EM}_{\mathrm{AC}}^{\alpha} = \frac{1}{|A_r|} \sum_{q_i \in A_r \cap A_g} \mathrm{EM}_{\mathrm{AC}}^{q_i}$ → 모델이 응답한 case 중 correctness 측정

    $\mathrm{EM}_{\mathrm{AC}}^{\beta} = \frac{1}{|A_g|} \sum_{q_i \in A_r \cap A_g} \mathrm{EM}_{\mathrm{AC}}^{q_i}$ → Answerable case 중 correctness 측정

    $A_r \cap A_g$ 이 summation의 delimiter가 되므로 unanswerable 상황에 응답을 많이 하면 penalty를 받게 됨

  - 최종 전체 dataset 차원 EM recall은 F1-score로 산출

$$\mathrm{EM}_{\mathrm{AC}}^{\mathrm{F1}} = \frac{2\ \mathrm{EM}_{\mathrm{AC}}^{\alpha} \cdot \mathrm{EM}_{\mathrm{AC}}^{\beta}}{\mathrm{EM}_{\mathrm{AC}}^{\alpha} + \mathrm{EM}_{\mathrm{AC}}^{\beta}}$$

TRUST-SCORE

- **Response Truthfulness: Grounded Refusal**

모델의 refusal (unanswerable), answerable question 대응 능력 파악
즉 answerable case중 응답을 생성한 비율, unanswerable case중 refusal 비율을 F1 관점으로 파악하겠다는 것
$\neg A_g$ 전체 case 중 unanswerable questions
$\neg A_r$ 모델이 응답하지 못한 questions
그래서 이상적인 상황은 $\neg A_g \cap \neg A_r = 1$ 이어야함. 즉 unanswerable questions에 대해서는 모두 refusal을 해야 함

$$\mathrm{F1_{RG}} = \frac{1}{2}(\mathrm{F1_{ref}} + \mathrm{F1_{ans}})$$

$$\mathrm{F1_{ref}} = \frac{2\mathrm{P_{ref}} \cdot \mathrm{R_{ref}}}{\mathrm{P_{ref}} + \mathrm{R_{ref}}} \qquad \mathrm{F1_{ans}} = \frac{2\mathrm{P_{ans}} \cdot \mathrm{R_{ans}}}{\mathrm{P_{ans}} + \mathrm{R_{ans}}}$$

$$\mathrm{P_{ref}} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_r|} \qquad \mathrm{P_{ans}} = \frac{|A_r \cap A_g|}{|A_r|}$$

$$\mathrm{R_{ref}} = \frac{|\neg A_r \cap \neg A_g|}{|\neg A_g|} \qquad \mathrm{R_{ans}} = \frac{|A_r \cap A_g|}{|A_g|}$$

* Over-responsiveness 측정 지표

$\mathrm{P_{ref}}$ LLM이 생성을 거부한 경우 중 실제 unanswerable의 비율

$\mathrm{R_{ref}}$ 실제 unanswerable 경우 중 LLM의 생성 거부 비율

* Excessive-refusal 측정 지표

$\mathrm{P_{ans}}$ LLM이 생성한 경우 중 answerable의 비율

$\mathrm{R_{ans}}$ 전체 answerable 경우 중 LLM의 생성 비율

TRUST-SCORE

- **Attribution Groundedness**

모델이 답변을 생성할 때 주어진 document를 잘 참조(cite)하면서 생성했는지 확인해야 함
- 모델이 생성한 답변을 statement 단위로 구분할 때 다음과 같이 구성될 수 있음

$s_i, \{c_{i,1}, c_{i,2}, ..., c_{i,j}\}$

→ 즉 모델이 생성한 문장 혹은 문단 s_i와 statement가 참조한 document citation numbering set $C_i = \{c_{i,1}, c_{i,2}, ..., c_{i,j}\}$ 으로 구성됨

* Citation Recall

- sample-wise citation recall

$$\mathrm{CR}^{s_i} = \phi(\{c_{i,1}, c_{i,2}, ..., c_{i,j}\}, s_i)$$

NLI 모델 ϕ를 활용하여 citation document set C_i이 s_i 를 entail하는지 확인

- dataset-wise citation recall

$$\mathrm{CR} = \frac{1}{|A_r|} \sum_{S \in A_r^s} \frac{1}{|S|} \sum_{s_i \in S} \mathrm{CR}^{s_i}$$

여기서 A_r^s는 모델이 생성한 response안에 포함된 statement의 집합

* Citation Precision

- sample-wise citation precision

$\mathrm{CP}^{c_j} = \phi(c_{i,j}, s_i)$ → 각 document $c_{\{i,j\}}$가 생성된 s_i 를 entail하는지 확인

$\mathrm{OR}\ \neg\phi(\{c_{i,k} | k \neq j\}, s_i)$

- dataset-wise citation precision

$$\mathrm{CP} = \frac{1}{|A_r|} \sum_{C \in A_r^c} \frac{1}{|C|} \sum_{c_j \in C} \mathrm{CP}^{c_j}$$

여기서 A_r^c는 모델이 생성한 response안에 포함된 citation의 집합

$$\mathrm{F1}_{\mathrm{CG}} = \frac{2\ \mathrm{CR} \cdot \mathrm{CP}}{\mathrm{CR}\ +\ \mathrm{CP}}$$

TRUST-SCORE

- **Responsivness**

  전체 case에 대하여 LLM의 응답 비율을 측정
  - 이상적으로 AR%는 dataset의 answerable distribution을 따라야 함

$$\mathrm{AR}\% = \frac{|A_r|}{|A_g| + \neg|A_g|}$$

TRUST-ALIGN DATASET

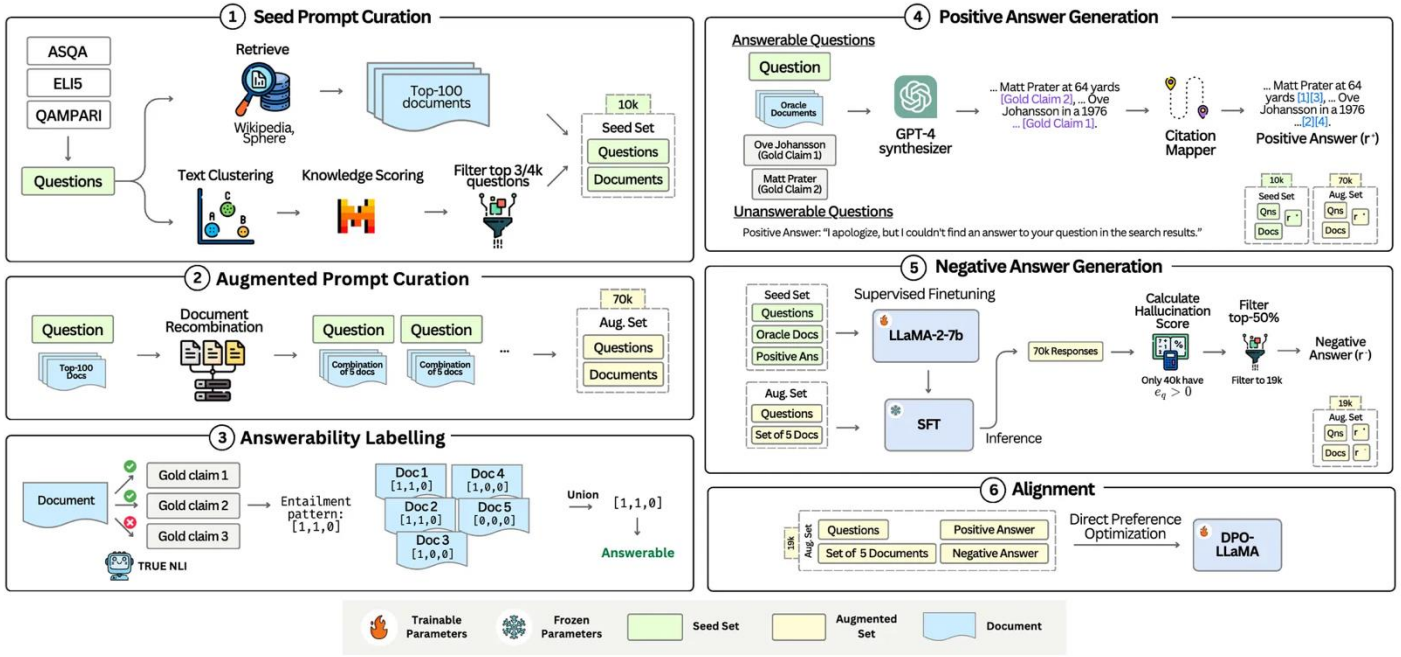


Figure 2: Overview of the TRUST-ALIGN. **Left**: The curation of both seed and augmented prompts (Q-D pairs) and an example of the answerability labeling process during the retrieval stage. **Right**: The response paired data generation process. First, we obtain positive answers and then select hard negative answers. Finally, we align our model via DPO.

RAG system하 LLM의 trustworthiness를 향상시키기 위한 alignment dataset인 TRUST-ALIGN DATASET을 구성

Experimental Setups

- **Datasets**

  Knowledge Attribution이 필요한 long-form QA dataset에 관하여 TRUST-SCORE TRUST-ALIGN dataset으로 학습한 모델과 baseline간의 비교실험 진행

  * In-domain
  - **ASQA**: 관점에 따라 여러 답변이 존재하는 ambiguous question에 관하여 답변 생성해야 하는 long-form QA dataset
  - **QAMPARI**: 질의에 응답하기 위해 여러 paragraph를 가지는 긴 문서에서 답변이 될 수 있는 모든 entity를 추출하는 QA dataset
  - **ELI5**: open-domain question에서 5살 아이도 이해할 수 있을 수준의 상세한 answer를 생성해야 하는 task

  * out-of-domain
  - **ExpertQA:** 32개의 전문분야 질문에 대하여 citation이 있는 answer를 생성해야하는 long-form QA task

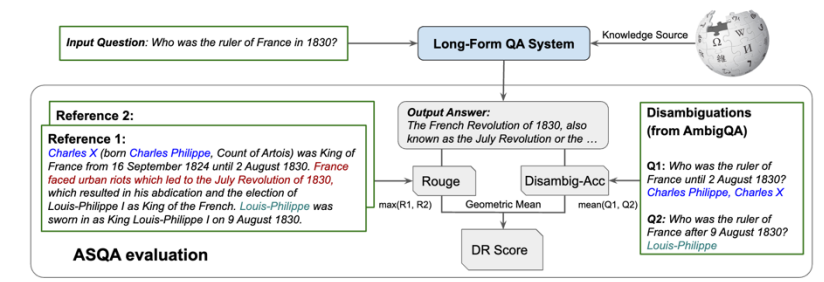


Figure 1: ASQA is an open-domain long-form QA dataset that focuses on answering ambiguous factoid questions. Input questions are sourced from AMBIGQA (Min et al., 2020). Long-form answers must be sufficient to answer disambiguated questions from AMBIGQA (short answers are marked in *blue* and *green*), and should introduce additional knowledge from Wikipedia (highlighted in *red*) to resolve ambiguity and clarify the relationship between different short answers. The DR score we propose combines ROUGE and Disambiguation-accuracy (that is, correctness) metrics, overcoming the issues with long-form QA evaluation outlined by Krishna et al. (2021).

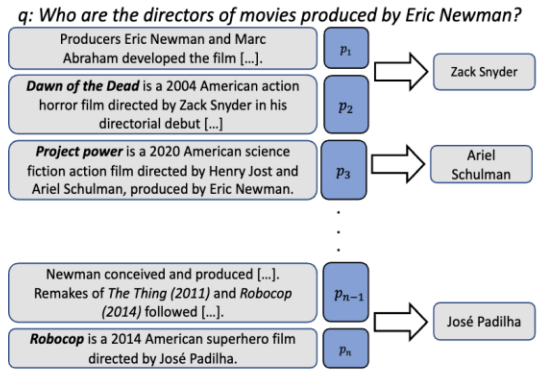


Figure 1: An example from QAMPARI with a generated question q, a subset of its evidence Wikipedia passages (left, p_i) and their corresponding answer.

Experimental Setups

- **Baselines**

Train-free setting

- ICL: Answer generation + inline citation을 수행하는 2-shot prompting을 주고 generation 실시
- PostCite: LLM이 retrieval 없이 바로 생성한 output에 대하여 GTR이 top-5 document에 대한 score를 활용하여 생성 output의 statement마다 citation matching 실시
- PostAttr: LLM이 retrieval 없이 바로 생성한 output에 대하여 각 statement별로 TRUE-NLI 모델의 score를 활용하여 생성 output의 statement마다 citation matching 실시

Trainable setting

- Self-RAG: LLM이 스스로 필요에 따라 retrieval을 요청하고 이를 활용한 생성 결과에 대하여 critique하는 방법으로 citation이 가미된 generation output 생성
- FRONT: LLM에게 grounding-guided generation training framework를 제안하는 연구. Grounding을 먼저 생성한 후에 citation이 달린 answer를 generation하는 연구 제안
- Trust-Align: LLaMA, Qwen, Phi 계열 모델에 대하여 TRUST-ALIGN dataset을 DPO 방식으로 학습

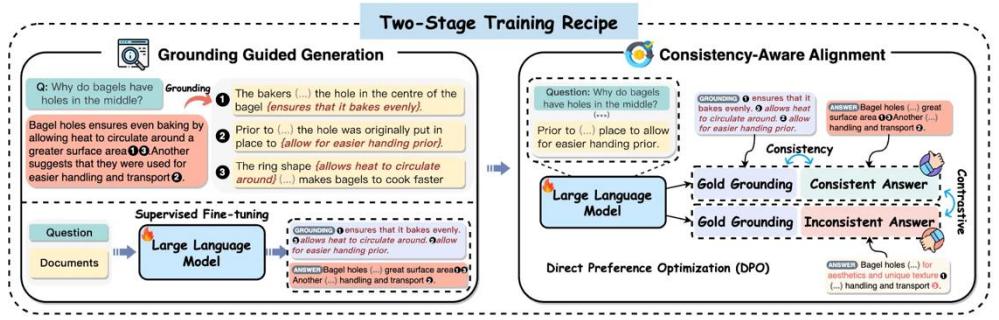


Figure 3: Overview of FRONT: The training recipe consists of two stages: grounding-guided generation and consistency-aware alignment. It enables LLMs to first generate precise grounding and subsequently guide the generation of attributed answers, thereby enhancing fine-grained attribution capability.

Result

- **Main result**

Model	Type	ASQA (610 answerable, 338 unanswerable)					QAMPARI (295 answerable, 705 unanswerable)					ELI5 (207 answerable, 793 unanswerable)				
		Resp.	Trustworthiness				Resp.	Trustworthiness				Resp.	Trustworthiness			
		AR (%)	Truthfulness		Att-Grd.	TRUST	AR (%)	Truthfulness		Att-Grd.	TRUST	AR (%)	Truthfulness		Att-Grd.	TRUST
			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$	
LLaMA-2 -7b	ICL	0.00	0.00	26.28	0.00	8.76	0.00	0.00	41.35	0.00	13.78	0.50	0.00	46.71	0.00	15.57
	PostCite	10.44	0.07	35.23	0.00	11.77	34.40	0.00	57.34	9.50	22.28	0.90	1.86	44.98	5.04	17.29
	PostAttr	10.44	0.07	35.23	0.00	11.77	34.40	0.00	57.34	3.78	20.37	0.90	1.86	44.98	0.00	15.61
	Self-RAG	100.00	45.19	39.15	63.49	49.28	96.00	6.81	28.23	19.95	18.33	73.50	14.94	40.20	13.80	22.98
	FRONT	100.00	60.47	39.15	68.86	56.16	100.00	17.27	22.78	24.26	21.44	100.00	21.66	17.15	52.72	30.51
	TRUST-ALIGN (DPO)	65.30	52.48	66.12	83.94	67.51	32.30	32.03	71.67	49.42	51.04	21.60	22.54	63.27	47.35	44.39
LLaMA-2 -13b	ICL	17.41	21.52	41.40	13.83	25.58	26.50	0.44	59.57	0.00	20.00	46.40	19.97	54.81	4.73	26.50
	PostCite	90.51	2.21	49.91	1.53	17.88	100.00	0.00	22.78	8.05	10.28	76.60	2.27	38.05	0.72	13.68
	PostAttr	90.51	2.21	49.91	0.17	17.43	100.00	0.00	22.78	2.95	8.58	76.60	2.27	38.05	0.09	13.47
	Self-RAG	100.00	48.52	39.15	69.79	52.49	72.70	2.71	48.58	26.91	26.07	22.10	12.77	58.68	24.54	32.00
LLaMA-3.2 -1b	ICL	60.23	35.95	50.94	9.96	32.28	19.20	6.32	52.64	0.38	19.78	88.40	12.87	27.10	5.23	15.07
	PostCite	43.57	0.59	50.22	0.24	17.02	41.20	0.32	49.79	1.61	17.24	18.40	2.04	50.88	1.02	17.98
	PostAttr	45.78	0.48	48.42	0.00	16.30	34.00	0.63	48.43	0.21	16.42	18.40	2.04	50.88	0.07	17.66
	FRONT	79.11	48.22	54.48	48.29	50.33	98.60	7.57	24.54	15.32	15.81	97.20	16.11	20.76	30.19	22.35
	TRUST-ALIGN (DPO)	41.67	38.64	58.61	79.35	58.87	20.00	27.22	67.92	49.42	48.19	9.60	13.20	59.35	48.21	40.25
LLaMA-3.2 -3b	ICL	1.27	2.04	27.98	53.95	27.99	34.10	16.06	59.65	12.87	29.53	21.90	18.55	55.56	30.70	34.94
	PostCite	47.26	31.03	56.59	22.99	36.87	39.60	6.34	55.22	6.83	22.80	92.80	18.12	25.14	4.44	15.90
	PostAttr	47.15	29.76	56.71	4.69	30.39	42.00	5.10	53.74	0.27	19.70	92.80	18.48	25.14	0.53	14.72
	FRONT	95.25	63.19	49.45	57.46	56.70	92.70	12.99	32.89	19.19	21.69	86.90	19.95	32.21	41.97	31.38
	TRUST-ALIGN (DPO)	77.85	59.82	66.38	84.21	70.14	48.20	29.13	70.85	45.65	48.54	17.50	18.33	62.79	55.87	45.66
LLaMA-3 -8b	ICL	1.48	3.01	28.58	86.50	39.36	3.90	5.92	48.60	20.24	24.92	0.00	0.00	44.23	0.00	14.74
	PostCite	77.53	32.98	53.31	28.01	38.10	87.00	6.10	34.52	8.42	16.35	62.00	20.80	45.88	8.06	24.91
	PostAttr	77.53	32.98	53.31	5.95	30.75	87.00	6.10	34.52	1.64	14.09	62.00	20.80	45.88	1.25	22.64
	FRONT	99.05	62.25	41.62	66.14	56.67	100.00	13.53	22.78	20.42	18.91	99.50	18.99	17.85	44.69	27.18
	TRUST-ALIGN (DPO)	56.43	53.94	65.49	88.26	69.23	22.40	35.35	70.73	58.77	54.95	15.50	20.81	63.57	50.24	44.87

TRUST-ALIGN boosts trustworthiness over baseline methods. As shown in Table 2 and Table 3, TRUST-ALIGNed models demonstrate substantial improvements on TRUST-SCORE over the baselines in 26 out of 27 model family and dataset configurations. Specifically, with LLaMA-3-8b, TRUST-ALIGN outperforms FRONT by 12.56% (ASQA), 36.04% (QAMPARI), and 17.69% (ELI5) on TRUST-SCORE. This suggests that TRUST-ALIGNed models are more capable of generating responses grounded in the documents.

• 제안한 TRUST-ALIGN 데이터를 활용하여 alignment를 진행할 때 전반적인 성능이 향상되어 truthfulness를 향상시킴

• 구체적으로 TRUST-ALIGN은 response 응답 비율을 낮추며 동시에 refusal 정확도를 타 baseline대비 크게 향상시켜 hallucinated answer generation을 크게 억제함

• 또한 TRUST-ALIGN은 citation F1을 향상에 크게 기여하여 grounding 능력을 향상시킴

→ LLM의 citation, refusal 능력을 향상시키는 TRUST-ALIGN dataset의 효과성 증명

•한편 답변의 정확성 correctness 측면에서는 양가적인 측면이 나타남.

이에 대하여 저자는 EM^{α}_{AC}, EM^{β}_{AC} 성능을 구분해서 분석을 실시

Result

- **Main result**

	Prompt	AR%	$\mathbf{EM_{reg}}$	$\mathbf{EM^{\alpha}_{AC}}$	$\mathbf{EM^{\beta}_{AC}}$	$\mathbf{EM^{F1}{}_{AC}}$	R_{ref}	P_{ref}	$F1_{ref}$	R_{ans}	P_{ans}	$F1_{ans}$	$\mathbf{F1_{RG}}$	CR	CP	$\mathbf{F1_{CG}}$	TRUST-SCORE
							LLaMA-3.2-3b										
ICL	R	1.27	0.63	52.78	1.04	2.04	99.41	35.90	52.75	1.64	83.33	3.22	27.98	53.47	54.44	53.95	27.99
PostCite	R	47.26	16.56	36.64	26.91	31.03	63.91	43.20	51.55	53.44	72.77	61.63	56.59	22.99	22.99	22.99	36.87
PostAttr	R	47.15	15.76	35.18	25.78	29.76	64.20	43.31	51.73	53.44	72.93	61.68	56.71	4.69	4.69	4.69	30.39
FRONT	R	95.25	40.68	52.94	78.37	63.19	10.95	82.22	19.32	98.69	66.67	79.58	49.45	60.04	55.09	57.46	56.70
SFT	R	68.04	23.75	47.89	50.64	49.23	51.48	57.43	54.29	78.85	74.57	76.65	65.47	80.23	71.52	75.63	63.44
TRUST-ALIGN (DPO)	R	77.85	31.81	54.63	66.09	59.82	42.31	68.10	52.19	89.02	73.58	80.56	66.38	85.00	83.43	84.21	70.14
TRUST-ALIGN (DPO-half)	R	62.45	23.28	50.74	49.24	49.98	56.80	53.93	55.33	73.11	75.34	74.21	64.77	83.95	77.08	80.37	65.04

•ASQA dataset LLaMA-3.2-3b 실험에서 TRUST-ALIGN은 $\mathrm{EM}^{\alpha}_{\mathrm{AC}}$(모델이 응답한 case 중 recall) 54.63으로 FRONT 52.94 대비 우수한 성능을 보임.

•그러나 $\mathrm{EM}^{\beta}_{\mathrm{AC}}$(전체 answerable case 중 recall)의 경우 낮은 성능을 보임

•한편 모델이 전체 응답해야 하는 question에 대해서 실제 응답한 비율을 나타내는 지표인 Rans는 TRUST-ALIGN이 89.02%, FRONT가 98.69%로 TRUST-ALIGN이 비교하여 더 낮았음

• 이러한 문제의 원인은 TRUST-ALIGN의 FRONT 대비 낮은 responsiveness 89.02% vs 92.25%에서 기인하는 것으로 추정

•다시 말하면 적게 생성하기 때문에 모델이 응답한 question의 개수가 줄어들어 **분모가 작아지는 α recall** potential은 높아짐
•반대로 전체 case를 대상으로 하는 β recall potential은 낮아지기 때문에 이러한 현상이 나타나는 것으로 예상

•저자는 이를 통해 TRUST-ALIGN**이 per-sample recall은 높을 가능성**이 존재한다고 주장

Result

- **Main result**

Table 3: Qwen2.5 and Phi3.5 families evaluated on the three datasets.

Model	Type	ASQA (610 answerable, 338 unanswerable)					QAMPARI (295 answerable, 705 unanswerable)					ELI5 (207 answerable, 793 unanswerable)				
		Resp.	Trustworthiness				Resp.	Trustworthiness				Resp.	Trustworthiness			
		AR (%)	Truthfullness		Att-Grd.	TRUST	AR (%)	Truthfullness		Att-Grd.	TRUST	AR (%)	Truthfullness		Att-Grd.	TRUST
			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$	
Qwen-2.5 -0.5b	ICL	29.85	20.96	47.19	0.35	22.83	11.40	2.45	50.67	0.00	17.71	82.30	13.73	33.14	0.37	15.75
	PostCite	46.10	8.55	50.84	8.23	22.54	17.00	0.67	52.51	5.72	19.63	89.80	9.87	27.10	4.10	13.69
	PostAttr	46.10	8.55	50.84	2.23	20.54	17.00	0.67	52.51	0.90	18.03	89.80	9.87	27.10	0.68	12.55
	FRONT	100.00	42.83	39.15	45.87	42.62	99.30	11.52	23.23	15.90	16.88	99.90	13.74	17.29	27.95	19.66
	TRUST-ALIGN (DPO)	71.84	50.59	61.28	52.40	54.76	17.90	15.76	61.84	29.73	35.78	21.70	13.68	60.79	22.72	32.40
Qwen-2.5 -1.5b	ICL	98.52	50.55	41.74	6.69	32.99	85.00	15.60	41.27	8.61	21.83	99.40	20.56	17.78	4.99	14.44
	PostCite	71.73	16.36	52.46	15.40	28.07	11.20	3.44	51.11	13.95	22.83	91.50	15.63	26.71	5.17	15.84
	PostAttr	71.73	16.36	52.46	4.45	24.42	11.20	3.44	51.11	1.07	18.54	91.50	15.63	26.71	0.62	14.32
	FRONT	99.26	57.74	41.36	55.70	51.60	98.80	16.05	24.45	11.60	17.37	99.90	19.57	17.29	37.70	24.85
	TRUST-ALIGN (DPO)	72.57	52.68	62.38	66.81	60.62	20.00	23.80	68.46	50.98	47.75	33.60	19.03	57.91	31.63	36.19
Qwen-2.5 -3b	ICL	27.43	37.72	51.36	51.72	46.93	22.30	23.17	63.27	41.20	42.55	68.80	29.12	46.31	34.34	36.59
	PostCite	8.76	9.58	35.30	10.94	18.61	0.10	0.00	41.31	0.00	13.77	49.70	21.73	48.49	7.56	25.93
	PostAttr	8.76	9.58	35.30	36.29	27.06	0.10	0.00	41.31	25.00	22.10	49.70	21.73	48.49	1.31	23.84
	FRONT	97.47	55.15	44.01	62.72	53.96	79.10	20.69	48.62	25.67	31.66	93.60	18.69	25.37	37.40	27.15
	TRUST-ALIGN (DPO)	49.47	55.19	63.76	78.64	65.86	48.10	35.69	70.31	45.64	50.55	13.50	22.52	64.38	42.01	42.97
Qwen-2.5 -7b	ICL	92.09	58.94	54.34	75.46	62.91	56.30	28.92	63.67	39.28	43.96	82.70	28.27	37.13	44.13	36.51
	PostCite	91.46	27.52	45.93	4.19	25.88	26.70	8.59	60.16	1.05	23.27	95.60	21.82	22.23	7.03	17.03
	PostAttr	91.46	27.52	45.93	17.92	30.46	26.70	8.59	60.16	13.55	27.43	95.60	21.82	22.23	0.96	15.00
	FRONT	86.39	64.58	60.08	58.27	60.98	84.70	17.02	42.85	24.48	28.12	57.60	28.27	54.14	56.61	46.34
	TRUST-ALIGN (DPO)	59.49	55.04	66.22	83.57	68.28	32.10	30.11	70.68	53.48	51.42	21.00	24.30	63.79	47.02	45.04
Phi3.5 -mini	ICL	63.19	50.24	51.95	42.64	48.28	70.20	11.91	43.90	12.26	22.69	81.50	27.59	37.17	30.14	31.63
	PostCite	23.10	14.98	41.38	9.40	21.92	76.90	3.57	42.36	4.49	16.81	84.50	20.50	30.81	4.67	18.66
	PostAttr	23.10	14.98	41.38	1.24	19.20	76.90	3.57	42.36	0.46	15.46	84.50	21.26	30.81	0.68	17.58
	FRONT	99.79	63.30	39.79	71.63	58.24	100.00	11.97	22.78	21.50	18.75	96.60	21.46	21.35	61.41	34.74
	TRUST-ALIGN (DPO)	66.56	52.23	64.20	85.36	67.26	30.10	36.42	73.95	53.40	54.59	24.90	23.39	67.62	47.42	46.14

TRUST-ALIGN generalizes across model families and sizes. Table 3 demonstrates that TRUST-ALIGN improves the models' TRUST-SCORE across various sizes and architectures. In small models like Qwen-2.5-0.5b, TRUST-ALIGN significantly outperforms ICL baselines, achieving notable gains in ASQA (22.83% → 54.76%). Similarly, for larger models such as Qwen-2.5-7b, TRUST-ALIGN delivers substantial improvements, as seen in ASQA (62.91% → 68.28%), highlighting its scalability. The largest gains are observed in smaller models; for example, Phi3.5-mini shows remarkable improvements over ICL: 18.98% (ASQA), 31.90% (QAMPARI), and 14.51% (ELI5).

Qwen 계열, Phi-3.5-mini에서도 일괄적인 성능 향상이 나타남

Table 4: Performance of models with only SFT applied as compared to TRUST-ALIGN models. Best values within each family are **bolded**).

Model	Type	ASQA (610 answerable, 338 unanswerable)					QAMPARI (295 answerable, 705 unanswerable)					ELI5 (207 answerable, 793 unanswerable)				
		Resp.	Trustworthiness				Resp.	Trustworthiness				Resp.	Trustworthiness			
		AR (%)	Truthfullness		Att-Grd.	TRUST	AR (%)	Truthfullness		Att-Grd.	TRUST	AR (%)	Truthfullness		Att-Grd.	TRUST
			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$			EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$	
LLaMA-2 -7b	SFT	80.17	**53.21**	63.43	79.61	65.42	31.60	**33.76**	71.13	46.37	50.42	29.50	21.58	**63.30**	39.59	41.49
	TRUST-ALIGN (DPO)	65.30	52.48	**66.12**	**83.94**	**67.51**	32.30	32.03	**71.67**	**49.42**	**51.04**	21.60	**22.54**	63.27	**47.35**	**44.39**
LLaMA-3.2 -1b	SFT	63.82	**45.61**	**63.91**	73.10	**60.87**	26.00	**27.98**	**68.20**	37.96	44.71	20.50	**14.56**	**63.93**	37.28	38.59
	TRUST-ALIGN (DPO)	41.67	38.64	58.61	**79.35**	58.87	20.00	27.22	67.92	**49.42**	**48.19**	9.60	13.20	59.35	**48.21**	**40.25**
LLaMA-3.2 -3b	SFT	68.04	49.23	65.47	75.63	63.44	27.60	28.09	70.22	38.03	45.45	14.70	15.92	62.59	53.33	43.95
	TRUST-ALIGN (DPO)	77.85	**59.82**	**66.38**	**84.21**	**70.14**	48.20	**29.13**	**70.85**	**45.65**	**48.54**	17.50	**18.33**	**62.79**	**55.87**	**45.66**
LLaMA-3 -8b	SFT	68.99	52.35	**66.06**	80.95	66.45	24.20	**33.85**	**71.11**	48.01	50.99	23.60	**22.57**	**65.06**	46.85	44.83
	TRUST-ALIGN (DPO)	56.43	**53.94**	65.49	**88.26**	**69.23**	22.40	35.35	70.73	**58.77**	**54.95**	15.50	20.81	63.57	**50.24**	**44.87**
Qwen-2.5 -0.5b	SFT	83.44	38.71	58.03	**57.47**	51.40	18.50	**16.02**	61.35	27.82	35.06	35.50	10.50	57.19	19.57	29.09
	TRUST-ALIGN (DPO)	71.84	**50.59**	**61.28**	52.40	**54.76**	17.90	15.76	**61.84**	**29.73**	**35.78**	21.70	**13.68**	**60.79**	**22.72**	**32.40**
Qwen-2.5 -1.5b	SFT	78.27	44.23	58.75	**71.08**	58.02	25.50	**23.89**	**69.66**	37.68	43.74	41.30	14.14	55.35	27.69	32.39
	TRUST-ALIGN (DPO)	72.57	**52.68**	**62.38**	66.81	**60.62**	20.00	23.80	68.46	**50.98**	**47.75**	33.60	**19.03**	**57.91**	**31.63**	**36.19**
Qwen-2.5 -3b	SFT	75.21	47.26	60.61	73.09	60.32	27.20	28.80	68.12	37.34	44.75	34.50	14.85	61.47	35.87	37.40
	TRUST-ALIGN (DPO)	49.47	**55.19**	**63.76**	**78.64**	**65.86**	48.10	**35.69**	**70.31**	**45.64**	**50.55**	13.50	**22.52**	**64.38**	**42.01**	**42.97**
Qwen-2.5 -7b	SFT	65.30	50.73	64.50	82.07	65.77	31.70	**33.58**	70.10	49.08	50.92	25.50	20.78	**64.25**	46.89	43.97
	TRUST-ALIGN (DPO)	59.49	**55.04**	**66.22**	**83.57**	**68.28**	32.10	30.11	**70.68**	**53.48**	**51.42**	21.00	**24.30**	63.79	**47.02**	**45.04**
Phi3.5 -mini	SFT	66.46	51.92	**64.34**	82.77	66.34	29.10	35.04	73.93	49.38	52.78	24.50	22.50	65.70	46.79	45.00
	TRUST-ALIGN (DPO)	66.56	**52.23**	64.20	**85.36**	**67.26**	30.10	**36.42**	**73.95**	**53.40**	**54.59**	24.90	**23.39**	**67.62**	**47.42**	**46.14**

Trust-Align DPO는 SFT대비 citation 능력 향상에 크게 기여함 (f1_CG)

다만 correctness (EM), Refusal (F1_RG) 경우에 대하여 SFT 대비 낮은 성능이 존재

Result

- Out-of-domain 실험 결과

Table 7: Generalization test results on ExpertQA using refusal prompting.

Model	Type	AR (%)	EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$	TRUST
LLaMA-2 -7b	ICL	0.51	0.00	41.01	9.52	16.84
	PostCite	5.62	4.85	44.27	5.23	18.12
	PostAttr	5.62	4.85	44.27	2.26	17.13
	FRONT	100	9.33	23.92	74.75	36.00
	TRUST-ALIGN (DPO)	20.01	25.03	67.91	62.46	**51.8**
LLaMA-3.2 -1b	ICL	90	21.55	32.83	9.04	21.14
	PostCite	30.84	5.48	49.1	2.67	19.08
	PostAttr	48.41	8.24	47.72	1.5	19.15
	FRONT	95.62	20.83	29.26	37.45	29.18
	TRUST-ALIGN (DPO)	15.44	20.32	64.87	62.1	**49.1**
LLaMA-3.2 -3b	ICL	58.74	33.5	51.21	38.37	41.03
	PostCite	82.85	25.68	38.11	5.29	23.03
	PostAttr	82.85	25.45	38.58	3.4	22.48
	FRONT	83.36	27.24	43.34	50.91	40.5
	TRUST-ALIGN (DPO)	7.24	11.72	56.93	78.35	**49.0**
LLaMA-3 -8b	ICL	0.65	2.82	42.5	69.46	38.26
	PostCite	15.68	14.06	50.08	7.09	23.74
	PostAttr	15.68	14.06	50.08	6.29	23.47
	FRONT	99.26	30.34	24.92	56.7	37.32
	TRUST-ALIGN (DPO)	16.41	27.36	67.07	70.11	**54.85**
GPT-3.5	ICL	59.47	36.65	56.39	63.93	52.32
GPT-4	ICL	72.20	41.32	52.91	69.83	**54.69**
GPT-4o	ICL	66.07	42.62	64.4	54.61	51.24
	TRUST-ALIGN (SFT)	36.84	28.85	71.68	61.98	**53.82**
Claude-3.5	ICL	73.95	11.68	51.91	10.7	24.76

Model	Type	AR (%)	EM^{F1}_{AC}	$F1_{RG}$	$F1_{CG}$	TRUST
Qwen-2.5 -0.5b	ICL	78.24	21.42	38.71	0.44	20.19
	PostCite	51.41	13.32	48.08	5.6	22.33
	PostAttr	51.41	13.32	48.08	1.49	20.96
	FRONT	99.86	18.27	24.05	34.62	25.65
	TRUST-ALIGN (DPO)	32.96	18.16	63.31	35.07	**38.85**
Qwen-2.5 -1.5b	ICL	98.34	30.67	26.09	6.89	21.22
	PostCite	62.19	22.22	48.66	16.92	29.27
	PostAttr	62.19	22.22	48.66	13.15	28.01
	FRONT	99.59	29.15	24.6	50.22	34.66
	TRUST-ALIGN (DPO)	30.2	25.06	68.38	51.44	**48.29**
Qwen-2.5 -3b	ICL	68.88	35.14	49.65	42.67	42.49
	PostCite	0.05	0	40.66	0	13.55
	PostAttr	0.05	0	40.66	0	13.55
	FRONT	95.48	25.67	29.86	44.48	33.34
	TRUST-ALIGN (DPO)	17.15	20.97	65.79	60.25	**49.0**
Qwen-2.5 -7b	ICL	84.56	36.33	42.28	56.09	44.9
	PostCite	42.14	25.58	54.9	13.77	31.42
	PostAttr	42.14	25.58	54.9	12.46	30.98
	FRONT	65.51	32.41	55.56	67.35	51.77
	TRUST-ALIGN (DPO)	24.99	25.57	69.16	62.7	**52.48**
Phi3.5 -mini	ICL	85.15	37.49	40.22	36.14	37.95
	PostCite	52.01	27.96	53.64	7.39	29.66
	PostAttr	52.01	27.96	53.64	5.7	29.1
	FRONT	97.37	28.19	27.5	65.82	40.5
	TRUST-ALIGN (DPO)	26.05	27.69	69.56	61.6	**52.95**

•TRUST-ALIGN이 다른 방법론 대비 Out-of-domain 상황에서도 높은 TRUST-SCORE를 보장

•한편 GPT 계열 모델의 경우 EM은 매우 높지만 refusal, citation 능력은 낮아짐. 즉 이는 해당 모델이 parametric knowledge에 기인하여 answer는 잘 생성하나 grounded generation 능력은 떨어진다고 할 수 있음

•Claude-3.5의 경우 EM 대비 refusal, citation 능력의 간극이 커짐

Discussion

- **LLM refusal and correctness tradeoff**
  - TRUST-ALIGN dataset으로 학습한 모델은 Refusal 생성 능력 향상이 가장 두드러졌음
  - 이러한 Refusal 능력의 향상은 hallucination 발생 위험성은 낮추나 correctness를 낮추고 excessive refusal case양상에 관여할 수 있음
    → RAG system이 구체적으로 어떠한 상황에서 답변을 할 수 없는지, 그리고 어떠한 경우에 excessive refusal이 발생하는지 연구 필요

- **Need for reference-free Truthfulness metric**
  - TRUST-SCORE의 기본 전제는 GT answer의 존재 + GT answer내 document 참조 부분 분해가 되어있다는 가정
  - 따라서 위와 같은 annotation이 없는 dataset + LLM응답 그 자체에 관해서 truthfulness 평가를 할 수 없는 metric임
    → Reference-free truthfulness metric 연구의 필요성

Thank you

atural Language
rocessing
Artificial Intelligence

고려대학교
KOREA UNIVERSITY