Evaluation Methods:
Assessment on the Model Generation

2024 여름방학세미나

NLP&AI 박찬희

Theme Selection

1. What do we focus on when we evaluate in different settings?

- 태스크 (요약, 대화, 번역, QA, …)
- 도메인 (의료, 과학, 수학, …)
- 언어 (한국어, 영어, …)
- 베이스라인 모델 (파라미터 크기, 모델 종류, …)

2. How do we express scores?

- Likert scale (1-5점 평가)
- Ranking (A/B test)
- Binary classification (0/1 label)
- Human annotation (reference-based, reference-free)
- Correlation (Pearson, Kendall's Tau, Spearman)

First Paper

ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems

Jon Saad-Falcon
Stanford University *
jonsaadfalcon@stanford.edu

Omar Khattab
Stanford University
okhattab@stanford.edu

Christopher Potts
Stanford University
cgpotts@stanford.edu

Matei Zaharia
Databricks and UC Berkeley
matei@databricks.com

NAACL 2024

Introduction

RAG 생성물 평가의 어려움

- QA, fact-checking, customer support 등 다양한 태스크 존재
- 도메인 지식을 반영한 전문가 평가 데이터 확보의 어려움

기존 RAGAS framework의 한계

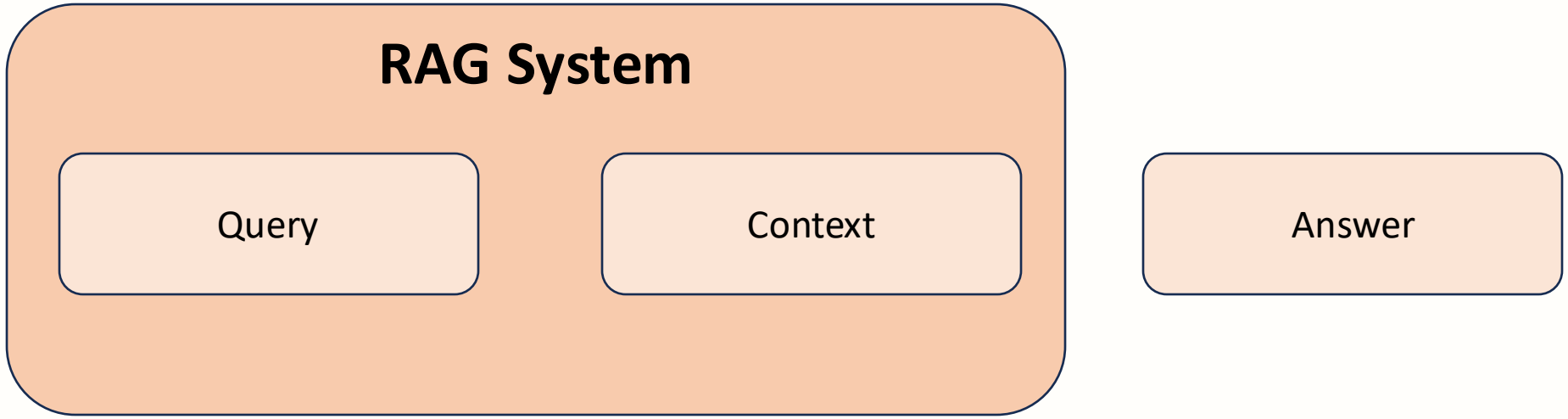
- 휴리스틱한 프롬프팅 기반 평가 방법으로 다양한 상황에 적용하기 어려움.
- GPT 평가 모델이 판단하는 점수의 퀄리티를 보장할 수 없음.

ARES의 차별점

- Prediction-Powered Inference (PPI)로 도출한 신뢰 구간 내 평가 결과를 제공
- Context relevance, answer faithfulness, answer relevance 3개 항목에 대해 평가
- 150개의 human annotation으로 데이터셋 전체에 대해 믿을만한 평가 지표 제공
- (모델 평가의 핵심은 결국 human evaluator를 대체할 수 있는가 (high correlation score)로 귀결)

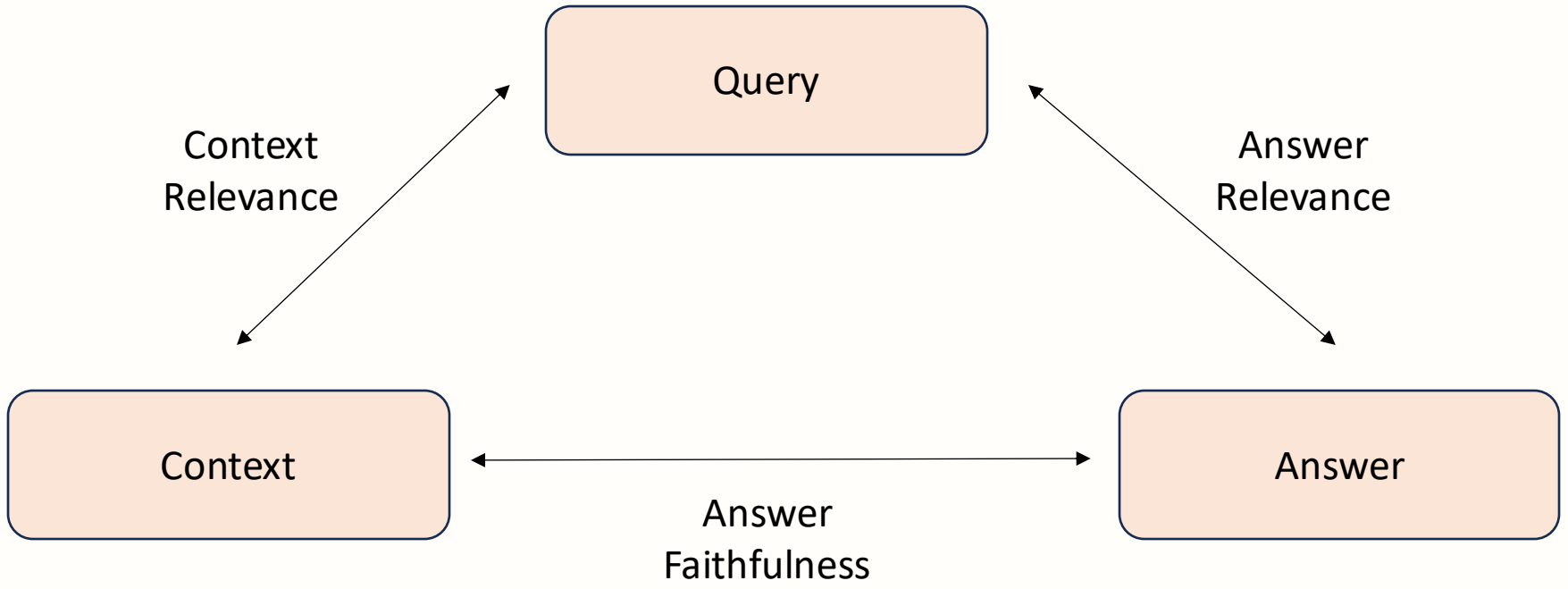
Introduction

- **Context relevance, answer faithfulness, answer relevance... Why these?**



Introduction

- **Context relevance, answer faithfulness, answer relevance... Why these?**



ARES

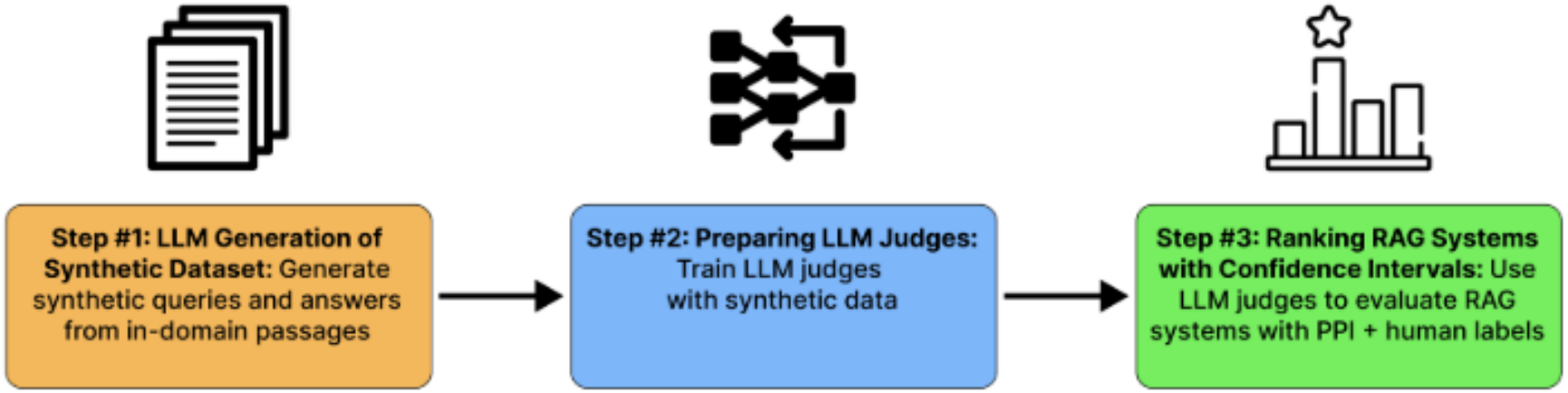


Figure 1: **Overview of ARES**: As inputs, the ARES pipeline requires an in-domain passage set, a human preference validation set of 150 annotated datapoints or more, and few-shot examples of in-domain queries and answers (five or more), which are used for prompting LLMs in synthetic data generation. To prepare our LLM judges for evaluation, we first generate synthetic queries and answers from the corpus passages. Using our generated training triples and a constrastive learning framework, we fine-tune an LLM to classify query–passage–answer triples in three different criteria: context relevance, answer faithfulness, and answer relevance. Finally, we use the LLM judges to score RAG systems and generate confidence bounds for the ranking using PPI and the human preference validation set.

ARES

Step 1. Synthetic Dataset Generation (FLAN-T5 XXL)

- Corpus passage (document)로부터 query-passage-answer triplet 생성
- few-shot prompting : FLAN-T5 XXL에게 few-shot 기반으로 query와 question을 순차적으로 생성
- Low-quality filtering : Retriever로 passage를 top result로 찾아낼 수 있는 경우는 모두 제외
- Positive/negative samples (same number of examples)

	Context Relevance	Answer Faithfulness	Answer Relevance
Weak Negative	Unrelated document	Randomly generated answer	
Strong Negative	Same doc, different passage (BM25 top10)	Generate contradictory answer	

ARES

Step 2. Preparing LLM(?) Judges (3 DeBERTa-v3-Large classifiers)

- **Context Relevance** : Is the passage returned relevant for answering the given query?
- **Answer Faithfulness** : Is the answer generated faithful to the retrieved passage, or does it contain hallucinated or extrapolated statements beyond the passage?
- **Answer Relevance** : Is the answer generated relevant given the query and retrieved passage?

Given query-passage-answer triplets,

positive answers : 1

negative examples : 0

ARES

Step 3. Ranking RAG Systems with Confidence Intervals

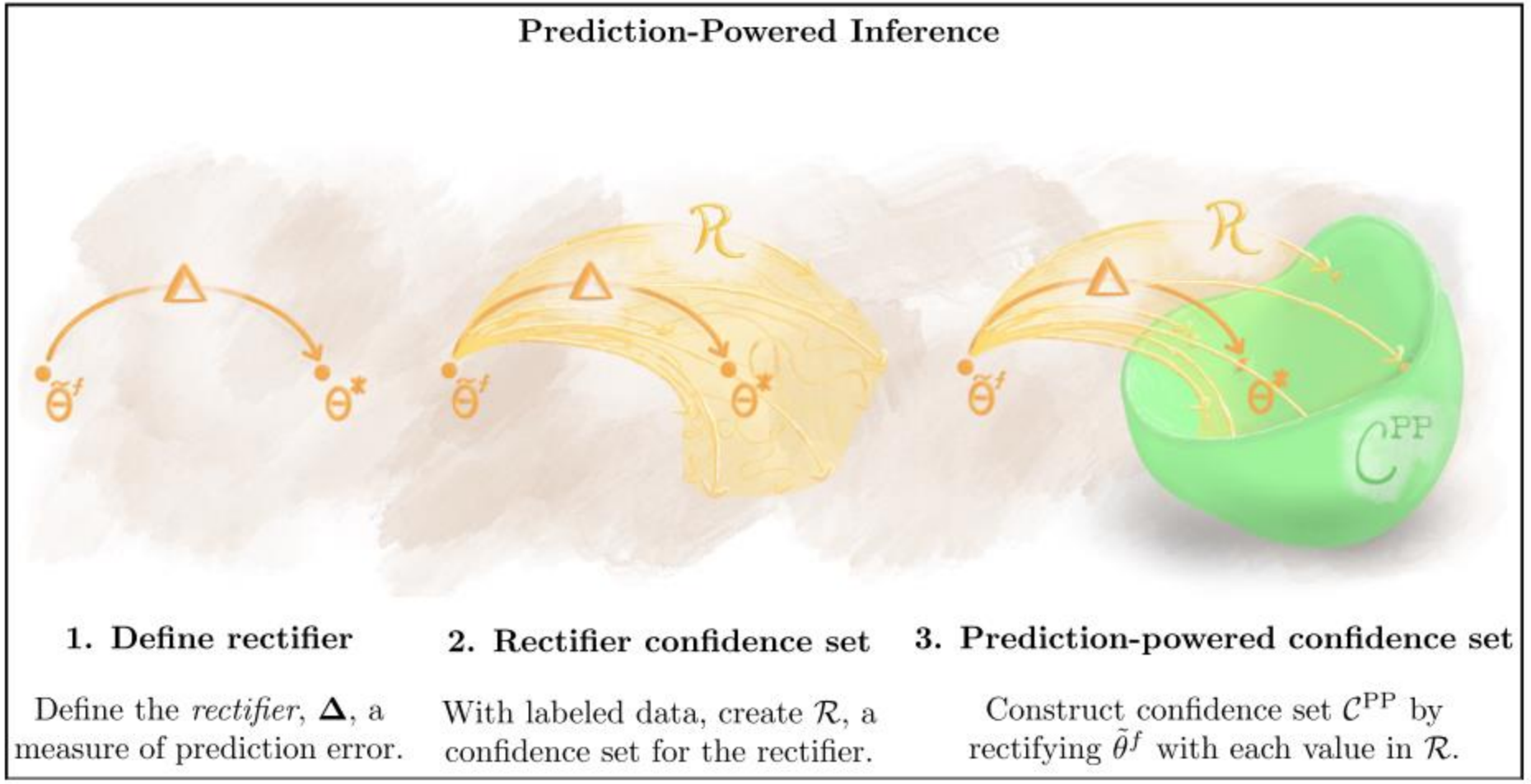
- 앞서 학습한 classifier로 그냥 평균 점수를 낼 수도 있지만, 합성 데이터 기반으로 학습했으므로 정확도가 의문
- **PPI** : 사람이 annotate한 validation set과 아주 근접한 신뢰 구간을 도출하는 통계 방법론
- **Rectifier function**

  "PPI uses the LLM judges on the human preference validation set to learn a rectifier function for constructing a confidence set of the ML model's performance, using each ML prediction in the larger non-annotated dataset."

  "Additionally, PPI allows us to estimate confidence intervals with a selected level of probability; for our experiments, we use a standard 95% alpha (probability) for our confidence interval.

- 앞선 과정에서 학습한 DeBERTa 모델의 응답을 PPI 알고리즘으로 '교정'한다.

PPI


Prediction-Powered Inference
$\mathcal{R}$
Δ
$\tilde{\theta}^f$
θ^*
C^{PP}
1. Define rectifier
Define the *rectifier*, Δ, a measure of prediction error.
2. Rectifier confidence set
With labeled data, create $\mathcal{R}$, a confidence set for the rectifier.
3. Prediction-powered confidence set
Construct confidence set $\mathcal{C}^{PP}$ by rectifying $\tilde{\theta}^f$ with each value in $\mathcal{R}$.

PPI

Machine prediction : larger, unlabeled, lots of data (#N, $(\widetilde{X}, \widetilde{Y})$)

Human-annotated data: smaller, labeled, gold-standard (#n, *(X, Y)*)

$\theta^* = \mathbb{E}[Y_i]$... classical한 의미로, 관측된 label에 대한 평균값 $\hat{\theta}^{\text{class}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

$$\hat{\theta}^{\text{PP}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} f(\widetilde{X}_i)}_{\tilde{\theta}^f} - \underbrace{\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)}_{\hat{\Delta}}.$$

만약 emperical rectifier인 $\hat{\Delta}$ 가 0에 근사한다면, $\hat{\theta}^{\text{PP}} \approx \frac{1}{N}\sum_{i=1}^{N} \widetilde{Y}_i$

라벨링 된 데이터 수 n보다 N이 훨씬 크므로, 위 근사식의 variance가 $\hat{\theta}^{\text{class}} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ 보다 작아짐.
이를 바탕으로 95% 신뢰 구간을 형성하면

$\underbrace{\hat{\theta}^{\text{PP}} \pm 1.96\sqrt{\frac{\hat{\sigma}^2_{f-Y}}{n} + \frac{\hat{\sigma}^2_f}{N}}}_{\text{prediction-powered interval}}$ 로 $\underbrace{\hat{\theta}^{\text{class}} \pm 1.96\sqrt{\frac{\hat{\sigma}^2_Y}{n}}}_{\text{classical interval}}$ 를 대체할 수 있음.

where $\hat{\sigma}^2_Y$, $\hat{\sigma}^2_{f-Y}$, and $\hat{\sigma}^2_f$ are the estimated variances of the Y_i, $f(X_i) - Y_i$, and $f(\widetilde{X}_i)$, respectively.

Experiment

Dataset & Metrics

- KILT : Natural Questions (NQ), HotpotQA, FEVER, Wizard of Wikipedia (WoW)
- SuperGLUE : MultiRC, ReCoRD
- 다양한 RAG 시스템이 존재하는 상황을 재현하기 위해 각 데이터셋에서 validation subset으로 70%~90% 사이 2.5% 구간으로 성공률을 할당, 이를 ranking 기준으로 삼음.
  "Since we know the success percentages of each dataset split, we know the appropriate ranking of each mock RAG system."
- **해석) 정답 순위를 알고 있기 때문에 ARES를 이용했을 때의 점수와 순위에 대한 성공률을 판단할 수 있음.**

- 실제 순위와 ARES가 평가한 순위간의 상관관계를 Kendall's Tau로 측정함.

$$\tau = \frac{(\#\text{ of concordant pairs}) - (\#\text{ of discordant pairs})}{\#\text{ of pairs total}}$$

Result & Analysis (C.R / A.R)

	ARES Ranking of Pseudo RAG Systems											
	NQ		HotpotQA		WoW		FEVER		MultiRC		ReCoRD	
	C.R	A.R.	C.R	A.R.	C.R	A.R.	C.R	A.R.	C.R	A.R.	C.R	A.R.
Kendall's Tau for Sampled Annotations	0.83	0.89	0.78	0.78	0.78	0.83	**0.89**	**0.89**	0.83	0.83	0.72	0.94
Kendall's Tau for RAGAS	0.89	0.89	**0.94**	0.89	0.94	0.94	0.72	0.61	0.83	**0.94**	**0.89**	0.44
Kendall's Tau for GPT-3.5 Judge	0.89	0.94	0.67	**0.94**	0.94	0.89	0.78	0.78	0.83	0.89	0.83	**0.94**
Kendall's Tau for ARES LLM Judge	0.89	**1.0**	0.89	**0.94**	0.94	**1.0**	0.83	0.72	**0.94**	0.83	0.78	0.83
Kendall's Tau for ARES	**0.94**	**1.0**	**0.94**	**0.94**	**1.0**	**1.0**	**0.89**	0.78	**0.94**	0.89	0.83	0.89
RAGAS Accuracy	31.4%	71.2%	17.2%	76.0%	36.4%	77.8%	23.7%	69.2%	16.1%	75.0%	15.0%	72.8%
GPT-3.5 Judge Accuracy	73.8%	95.5%	75.3%	71.6%	84.3%	85.2%	60.4%	59.6%	72.4%	60.3%	81.0%	65.8%
ARES Accuracy	79.3%	97.2%	92.3%	81.3%	85.7%	96.1%	88.4%	78.5%	85.8%	82.7%	67.8%	92.3%

Result & Analysis (A.F)

	WoW	CNN / DM
ARES Split Prediction	0.478	0.835
Correct Positive/Negative Split	0.458	0.859
ARES Judge Accuracy	62.5%	84.0%
Evaluation Set Size	707	510
Human Preference Data Size	200	200

Table 2: ARES Results on the AIS benchmark

AIS attribution benchmark

- 모델이 faithful한 답변과 hallucinated된 답변을 구분할 수 있는지에 대한 실험
- 각 문항은 passage가 answer에 faithful한지, non-attributed인지 구별하는 문항

RAG Systems Ranking

RAG Systems

Retriever : BM25, OpenAI Ada, ColBERTv2

LLM : MPT-7b-Instruct, GPT-3.5-Turbo, GPT-4

+ Facebook RAG (DPR + BART seq2seq)

Result

- 95% 이상 신뢰구간으로 정답 예측
- PPI confidence interval은 C.R의 경우 7.4, A.R의 경우 6.1

	ARES Ranking of Real RAG Systems					
	NQ		WoW		FEVER	
	C.R.	A.R.	C.R.	A.R.	C.R.	A.R.
Kendall's Tau for Sampled Annotations	0.73	0.78	0.73	0.73	0.73	0.82
Kendall's Tau for RAGAS	0.82	0.82	0.73	0.82	0.73	0.87
Kendall's Tau for GPT-3.5 Judge	0.82	0.87	0.82	0.82	0.64	0.87
Kendall's Tau for ARES LLM Judge	0.91	**0.96**	**0.91**	**1.0**	0.73	0.87
Kendall's Tau for ARES	**1.0**	**0.96**	**0.91**	**1.0**	**0.82**	**1.0**
RAGAS Accuracy	35.9%	68.2%	44.4%	80.1%	21.4%	75.9%
GPT-3.5 Accuracy	80.5%	91.2%	81.2%	83.5%	61.3%	54.5%
ARES Accuracy	85.6%	93.3%	84.5%	88.2%	70.4%	84.0%

RAG Systems Ranking

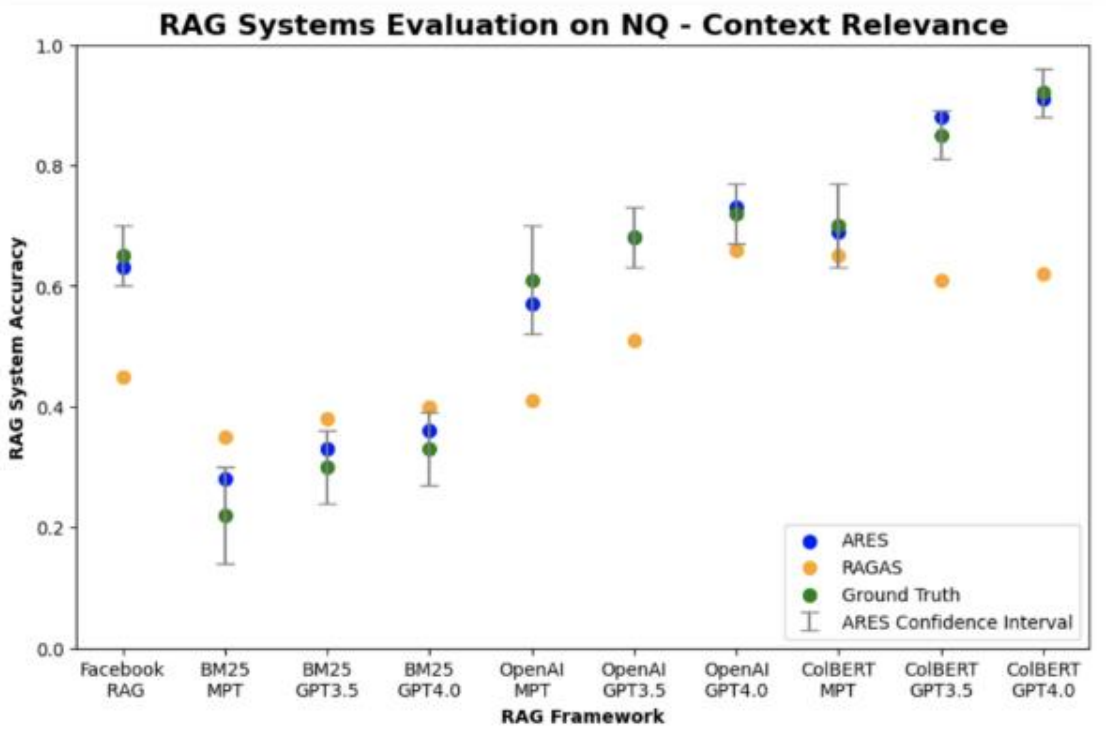


Figure 2: **RAG Systems Evaluation on NQ - Context Relevance**

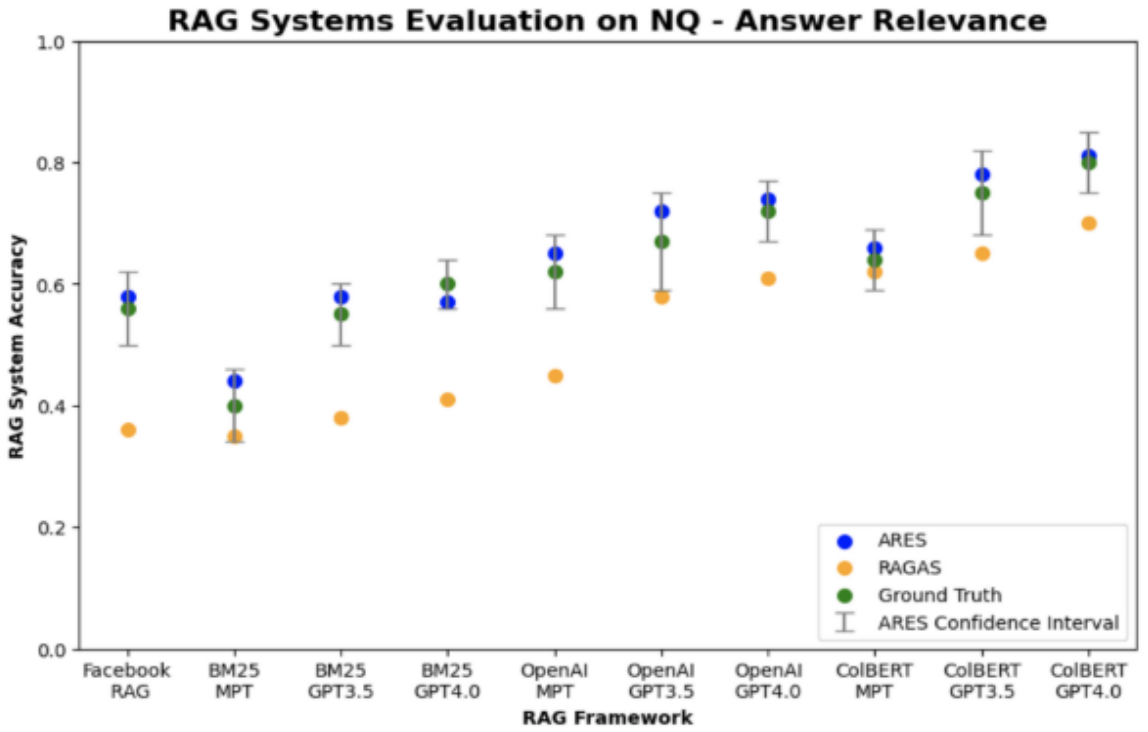


Figure 3: **RAG Systems Evaluation on NQ - Answer Relevance**

Cross-Domain Usage

	ARES Cross-Domain Ranking of Pseudo RAG Systems											
	NQ to FEVER		FEVER to NQ		NQ to MultiRC		MultiRC to NQ		NQ to ReCoRD		ReCoRD to NQ	
	C.R.	A.R.	C.R.	A.R.	C.R.	A.R.	C.R.	A.R.	C.R.	A.R.	C.R.	A.R.
Kendall's Tau	0.89	0.89	1.0	0.83	0.94	0.89	1.0	0.89	0.78	0.89	0.89	0.94
Kendall's Tau of In-Domain LLM Judge	0.89	0.78	0.94	1.0	0.94	0.89	0.94	1.0	0.83	0.89	0.94	1.0
Average PPI Range	8.7%	7.2%	6.5%	11.5%	10.2%	11.3%	11.9%	11.5%	10.5%	10.1%	9.7%	6.2%
Accuracy on RAG Evaluation Sets	92.4%	28.4%	85.7%	22.6%	81.5%	92.1%	87.6%	80.2%	29.1%	81.2%	80.1%	92.1%

ARES Judge 모델이 얼마나 강건한가?

NQ/FEVER : query type 차이

NQ/MultiRC : document type 차이

NQ/ReCoRD : query/document type 차이

- PPI로 out-of-domain인 경우에도 어느 정도 accuracy를 선방했다.

총평

1.낮은 GPT 의존도
- GPTScore, G-Eval 등 대세인 LLM 기반 평가 모델보다 강건하면서도 오픈소스 베이스라인 모델 선정.
- 프레임워크 전체에서 GPT를 한 번도 사용하지 않음.

2.적은 human eval, 높은 효율
- 실제 서비스를 개발하는 입장에서 참조 데이터를 150개만 구해도 된다는 점이 굉장히 매력적일 것.

1.이진분류 모델
- 각 항목의 반영 여부를 0/1으로 설정하여 단순화한 만큼 평가의 detail이 아쉬움.

2.Human error를 고려하지 않음
- 사람의 판단을 gold standard, ground truth로 선언한 만큼, human annotated 데이터의 품질이 중요

3. 다양한 태스크에 모델을 개별 학습
- 실험에서도 각 벤치마크마다 별개의 모델을 학습하여 실험했음.
- 태스크가 다양할수록 증강 및 모델 학습에 대한 부담이 커짐.

Second Paper

LLM Comparative Assessment: Zero-shot NLG Evaluation through Pairwise Comparisons using Large Language Models

Adian Liusie, Potsawee Manakul, Mark J. F. Gales
ALTA Institute, Department of Engineering, University of Cambridge
al826@cam.ac.uk, pm574@cam.ac.uk, mjfg@eng.cam.ac.uk

EACL 2024

Introduction

Human Eval의 한계

- 다양한 NLG 과제에서 사람의 평가를 gold standard로 삼고 있는데, 이는 비싸고 품이 많이 든다.
- 이를 대체하기 위한 자동화 평가 방법론은 task에 따라 다양한 제약을 받는다. (DialoGPT, SelfcheckGPT, Mqag)
- LLM의 창발 능력에 따른 prompt 기반 평가 방법론은 모델 크기의 제약이 존재한다.

사람의 평가 방식

- 점수를 개별적으로 내기보다 두 선택지를 비교하는 것이 더 낫다.
- 이를 확장, 다양한 태스크와 텍스트 관계에 적용할 수 있는 간단한 평가 모델을 만들자.

Contribution

1. NLG 평가에 포괄적인 pairwise 비교 분석을 적용한 첫 연구
2. 중간 크기 LLM에 대해 prompt-scoring보다 비교 분석 방법이 더 나음을 증명함
3. 위치 편향이 비교 분석에 영향을 줌을 밝히고, 이를 해소하기 위한 방법론을 제안함

Comparative Assessment

$y_{ij} \in \{0, 1\}$: x_i와 x_j의 랭크를 비교

$$p_{ij} = P(y_{ij}|x_i, x_j, d) \quad (1)$$

p_ij : x_i가 x_j보다 나은 답변일 확률 (d : 컨텍스트)

$$\hat{y}_{ij} = \begin{cases} 1, & \text{if } p_{ij} > 0.5 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

위 (1) 식을 hard decision으로 판단하는 함수

N개의 생성 문장에 대해서 R=N(N-1)번의 순위를 계산하면 전체 순위를 구할 수 있음
과정의 효율화를 위해 random, no-repeat, symmetric 3개의 그룹으로 평가 전략을 채택

- Random : 모든 경우의 수를 임의로 비교
- No-repeat : (x_i, x_j)를 비교했다면 (x_j, x_i)는 선택하지 않음
- Symmetric : (x_i, x_j)를 비교했다면 (x_j, x_i)도 비교함

위 과정을 거친 뒤 **win-loss ratio**로 응답 점수를 기록

$$\hat{s}_i = \frac{\#\text{wins of } x_i}{\#\text{comparisons involving } x_i} \quad (6)$$

Comparative Assessment

Prompt-Based Classifier

$$p_{ij} = \frac{P_\theta(w_i|\mathcal{P})}{P_\theta(w_i|\mathcal{P}) + P_\theta(w_j|\mathcal{P})} \quad (3)$$

: w는 class decision을 나타내는 라벨, P는 프롬프트

Text Generation

$$p_{ij} = \frac{1}{K}\sum_{k=1}^{K} f(\tilde{w}^{(k)}) \quad (5)$$

: k개의 토큰과 프롬프트가 주어졌을 때
라벨 토큰을 생성할 확률

$\tilde{y}_{ij}$: 예측하고자 하는 결과

$\hat{y}_{ij}$: i와 j 의 순서를 바꾸었을 때의 결과

$$P(A) = \frac{\sum_{i,j\in\mathcal{C}} \hat{y}_{ij}}{|\mathcal{C}|} \quad P(B) = \frac{\sum_{i,j\in\mathcal{C}} \tilde{y}_{ij}}{|\mathcal{C}|} \quad (8)$$

위 식에서 두 y의 값이 같다면 P(A)와 P(B)도 같아야 한다.

Comparative Assessment

Debaising Method

이상적인 경우라면, 순서를 바꾼 순위 결과가 동일해야 하는데, LLM은 처음 나온 답변을 선호하는 등의 위치 편향을 보임. 이는 거대 LLM일수록 더욱 심하게 드러남.

이에 따라 A를 선택할 선험 확률 P(A)와 B의 선험 확률 P(B)를 비교, 아래 식을 통해 debiasing을 진행함. (이상적인 경우라면 P(A) = P(B) = 0.5)

$$\hat{p}_{ij} = \frac{\alpha \cdot p_{ij}}{\alpha \cdot p_{ij} + (1 - p_{ij})} \quad (9)$$

예시) p_ij = **0.33**, alpha = 2

P^ij (system prob) = 0.66 / 0.66+0.66 = **0.5**

$$\hat{y}_{ij} = \begin{cases} 1, & \text{if } p_{ij} > \tau \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

이때, 보정된 확률 임계값 T는 일반적인 0.5가 아니라 **0.33**이 되어야 한다.

즉, alpha를 통해 P(A)=P(B)=0.5인 상황을 맞춰주고, 이에 맞게 임계값 Tau를 조정해준다.

Experiment

Dataset

- SummEval : 100 passages, 16 summaries
- Podcast : 179 podcasts, 15 abstractive summaries
- TopicalChat : 60 dialouges, 6 responses
- WebNLG : 223 semantic triplets, 8 responses

Base LLM

- FlanT5 220M, 770M, 3B, 11B
- LLaMA2-chat 3B, 13B

Experiment

Evaluation Metrics (for comparison)

- BLEU, ROUGE, BERTScore
- Bespoke (task-specific) : UniEval, QuestEval, MQAG, Longformer-SFT
- Zero-shot : GPTScore, Prompt Scoring (FlanT5, LlaMA2), G-Eval (prompting)

*베이스라인 모델은 comparative assessment, prompt-scoring을 모두 적용함.

Experiment

1. Summary Assessment

Approach	COH	CON	FLU	REL
Baselines (§4.3)				
BERTScore (w/ Ref)	25.9	19.7	23.7	34.7
QuestEval	18.2	30.6	22.8	26.8
MQAG	17.0	28.8	19.3	16.6
UniEval (single-best)	54.6	47.2	43.3	46.3
UniEval (continual)	57.5	44.6	44.9	42.6
GPTScore FlanT5-3B	47.0	43.6	42.1	34.4
GPTScore FlanT5-11B	45.6	43.8	42.4	34.3
GPTScore GPT3	40.1	47.5	41.0	34.3
ChatGPT scoring†	45.1	43.2	38.0	43.9
Prompt Scoring (§4.3.2)				
FlanT5-220M	4.0	-0.2	0.2	2.8
FlanT5-770M	-3.6	-1.6	-1.5	-0.0
FlanT5-3B	**14.5**	**19.8**	**3.9**	**15.2**
FlanT5-11B	0.7	11.2	3.2	5.7
Llama2-chat-7B	8.6	9.0	1.8	7.8
Llama2-chat-13B	9.9	6.9	1.2	9.2
G-Eval (§4.3.2)				
FlanT5-220M	3.6	0.6	2.7	8.0
FlanT5-770M	8.5	7.0	15.3	24.1
FlanT5-3B	10.5	29.1	9.8	23.8
FlanT5-11B	19.2	29.3	20.7	35.8
Llama2-chat-7B	28.2	29.4	**23.0**	27.4
Llama2-chat-13B	**53.2**	**33.7**	16.5	**38.3**
Comparative Assessment (§3)				
FlanT5-220M	4.0	-0.2	0.2	2.8
FlanT5-770M	29.8	26.3	20.6	35.1
FlanT5-3B	**51.2**	**47.1**	**32.5**	44.8
FlanT5-11B	44.2	37.2	30.2	43.4
Llama2-chat-7B	27.9	24.6	20.2	35.6
Llama2-chat-13B	40.9	39.9	30.8	**45.3**

Table 1: Spearman correlation coefficient for **SummEval**, averaged over both prompts per system (for prompt-scoring and comparative). †ChatGPT performance is quoted from Wang et al. (2023), which use more detailed scoring prompts.

- Prompt Scoring could be an emergent ability…
- GPTScore : Text Generation
- G-Eval : Prompt-Scoring (with task specific prompts)

Experiment

2. Podcast Assessment

Approach	System-lvl	Summary-lvl
Baselines (§4.3)		
BERTScore (w/ Ref)	73.9	25.1
UniEval (continual)	42.0	22.8
QuestEval	42.5	20.4
MQAG	77.9	12.6
Longformer-SFT	89.6	19.6
Prompt Scoring (§4.3.2)		
Llama2-chat-7B	88.5	2.6
Llama2-chat-13B	80.0	25.3
Comparative Assessment (§3)		
Llama2-chat-7B	88.2	37.4
Llama2-chat-13B	97.1	45.5

Table 2: Spearman correlation coefficient for **Podcast**.

- FLAN-T5는 토큰 수 초과로 제외
- Summary-lvl에서 coefficient가 상당히 낮게 나옴.
  (lack of granularity)

3. Dialogue Assessment

Approach	COH	CNT	ENG	NAT
Baselines (§4.3)				
UniEval (single-best)	60.7	-	59.6	54.7
UniEval (continual)	61.3	-	60.5	44.4
GPTScore GPT3	56.9	32.9	49.6	52.4
ChatGPT scoring†	54.7	57.7	37.9	58.0
Prompt Scoring (§4.3.2)				
FlanT5-220M	-2.2	0.2	-8.4	2.1
FlanT5-770M	3.7	3.1	-4.3	3.8
FlanT5-3B	**31.9**	**28.8**	17.4	**23.7**
FlanT5-11B	15.3	8.0	4.3	24.3
Llama2-chat-7B	16.4	17.0	20.6	21.4
Llama2-chat-13B	21.7	19.9	**31.4**	23.2
Comparative Assessment (§3)				
FlanT5-220M	-0.3	8.2	-10.5	2.2
FlanT5-770M	38.5	36.3	25.3	35.3
FlanT5-3B	49.4	**49.4**	37.3	47.4
FlanT5-11B	**54.3**	42.2	54.7	**54.2**
Llama2-chat-7B	28.9	33.7	36.1	30.3
Llama2-chat-13B	32.4	43.2	**55.5**	33.5

Table 3: Spearman correlation coefficient for **TopicalChat**.
†ChatGPT is prompted using our prompt-scoring prompts.

- Summary 결과와 유사

Experiment

4. Data-to-Text Assessment

- Triplet의 이해 능력도 LLM의 emergent ability일 것
- 3B/11-13B 모델에서 grammar와 fluency는 큰 차이가 없으나, semantic understanding에서 큰 차이 발생

Approach	FLU	GRA	SEM
Baselines (§4.3)			
BLEU	36.3	34.7	50.3
METEOR	44.3	42.9	62.7
NLI Model*	-	-	63.7
UniEval (continual)	21.7	16.3	-
Prompt Scoring (§4.3.2)			
FlanT5-220M	18.5	17.4	8.0
FlanT5-770M	14.5	13.6	17.1
FlanT5-3B	**30.8**	**32.7**	**38.5**
FlanT5-11B	-0.7	6.9	20.8
Llama2-chat-7B	3.8	2.4	17.0
Llama2-chat-13B	1.8	0.5	5.6
Comparative Assessment (§3)			
FlanT5-220M	-13.6	-17.9	0.1
FlanT5-770M	36.2	35.2	11.4
FlanT5-3B	40.6	41.4	12.8
FlanT5-11B	41.4	44.8	52.4
Llama2-chat-7B	22.9	37.8	-5.3
Llama2-chat-13B	**44.9**	**45.1**	**53.5**

Table 4: Spearman correlation coefficient for **WebNLG**. *Quoted from the NLI method with the backoff template in Dušek and Kasner (2020).

Positional Bias

- 일부 환경에서는 처음 선택지를 선택할 확률이 80%에 달할 정도로 위치 편향이 심하다.
- 큰 모델일수록 위치 편향에 취약하다.

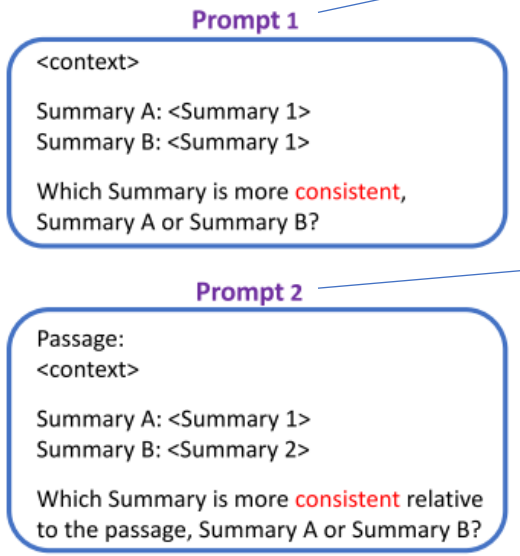


Figure 2: Comparative prompt template 1 and 2. When assessing different attributes, only the attribute is changed (e.g., consistent → engaging) and for response assessment, the word 'summary' is replaced with 'response'.

System	Prompt	COH	CON	FLU	REL
FlanT5 3B	1	0.37	0.46	0.39	0.41
	2	0.43	0.47	0.40	0.44
FlanT5 7B	1	0.18	0.20	0.13	0.23
	2	0.24	0.24	0.17	0.26
Llama2-chat 7B	1	0.41	0.17	0.26	0.18
	2	0.68	0.56	0.48	0.45
Llama2-chat 13B	1	0.31	0.37	0.18	0.32
	2	0.29	0.30	0.19	0.26

Table 5: Positional bias $P(A)$ for both prompt templates, for various systems in the comparative setup on SummEval.

Positional Bias

Debiased Result

System	Debias	SummEval				TopicalChat				WebNLG		Avg.
		COH	CON	FLU	REL	COH	CNT	ENG	NAT	FLU	GRA	
FlanT5-3B	✗	51.2	47.1	32.5	44.8	49.4	49.4	37.3	47.4	41.0	41.8	44.2
	✓	51.8	46.9	33.0	45.3	49.6	50.2	38.0	46.3	40.7	42.3	44.4
FlanT5-11B	✗	44.2	37.2	30.2	43.4	54.3	42.2	54.7	54.2	41.4	44.8	44.7
	✓	45.3	39.7	30.7	44.7	57.2	59.5	59.5	58.8	44.5	44.6	48.5
Llama2-chat-7B	✗	29.4	24.6	19.7	35.2	28.2	33.1	36.3	28.7	22.9	37.8	29.6
	✓	28.8	24.8	19.7	35.5	29.1	34.5	39.7	28.5	24.3	37.1	30.2
Llama2-chat-13B	✗	40.9	39.9	30.8	45.3	32.4	43.2	55.5	33.5	44.9	45.1	41.2
	✓	42.8	40.3	31.9	47.1	32.5	44.5	56.9	38.4	45.9	43.7	42.4

Table 6: Spearman correlation coefficient on different aspects of the NLG evaluation tasks, averaged over all prompts considered, using all pairs and ordering considered (i.e. full matrix comparisons).

Positional Bias

Debiased Result

System	Debias	COH	CON	FLU	REL
FlanT5-3B	✗	68.6	82.0	68.2	67.2
	✓	69.8	82.1	68.8	67.8
FlanT5-11B	✗	61.6	70.3	60.3	63.3
	✓	66.2	76.7	65.9	67.4
Llama2-chat-7B	✗	59.6	63.8	59.6	61.0
	✓	60.3	65.7	60.4	63.1
Llama2-chat-13B	✗	62.6	75.4	61.1	65.4
	✓	65.8	76.9	67.2	68.5

Table 7: Accuracy of the comparative systems, at a comparison level, for SummEval.

Computation Cost

Self-Consistency

SummEval의 경우, passage마다 총 240번의 비교가 이루어져야 한다. ((16 x 15) x 100 passages)
이를 줄였을 때의 결과를 최종 결과와 비교하는 실험.

	2	3	4	6	8	12	16
Final	84.0	88.3	90.7	93.7	95.5	98.0	100
Gold	68.0	69.1	69.7	70.3	70.6	70.8	70.9

Table 8: Accuracy when using fewer systems with respect to final rankings (using all 16 systems) and the ground truth labels. Results shown for Summeval COH using FlanT5-xl.

2의 경우, passage마다 2x1개씩 선택하여 총 200개 생성문을 비교 ((2 x 1) x 100 passages)
3의 경우, passage마다 3x2개씩 선택하여 총 600개 생성문을 비교 ((3 x 2) x 100 passages)
…

Computation Cost

Subset of Comparisons

이번에는 비교하는 수를 고정해놓고, sample을 선정하는 전략을 no-repeat, random, symmetric으로 나누어 비교함.
성능 지표인 Spearman 계수의 측면에서, 동일 비교수 대비 차이는 크지 않다.
단, debias 해준 결과의 성능은 유의미하게 상승함.
특히 비교 수가 적을 때 debias의 효과가 더 크다.

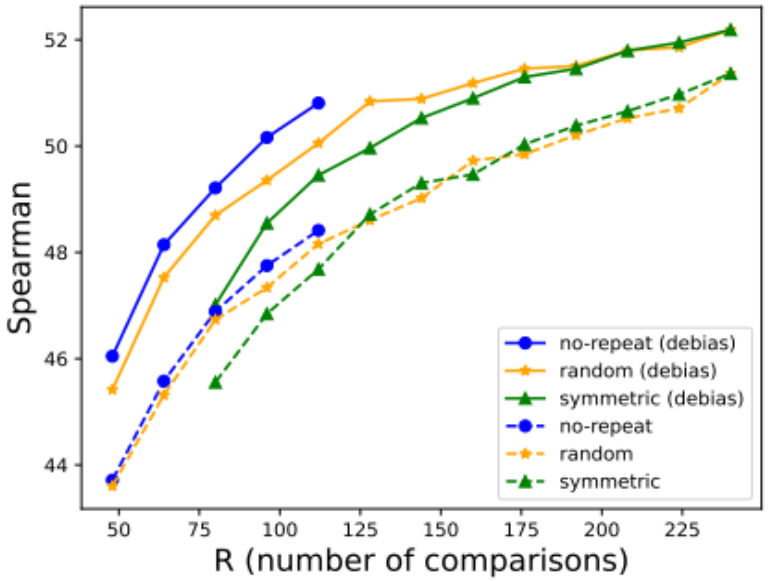


Figure 4: FlanT5-3B performance for SummEval COH when a subset of the comparisons are selected by either random, no-repeat or symmetric (as described in §3.4). For no-repeat, each pair is compared once, hence has a smaller maximum R.

총평

1.위치 편향 문제의 지적과 해소

- 단순 A/B 선택 기반 ranking 실험으로 끝났다면 EACL에 등재되기 어려웠을 것

2.구현 난이도

- 간단한 프롬프팅 기반 설계와 작은 모델 사이즈로 구현이 가능하다는 점
- 다양한 태스크에 대해서 별도의 모델 학습 없이 베이스라인 모델의 프롬프트 이해 능력을 기반으로 평가 방식을 구현한 점

1.너무하는 연산량 O(N^2)

- 본 논문에서 제안하는 성능을 온전히 구현하려면 비교해야 하는 경우의 수가 너무 많음.
- 단, 1대1 비교라면 scoring보다 나은 대안일 수 있음

2.적은 비교 데이터

- 위의 문제 때문에 task별로 100~200개의 passage에 대해서만 비교한 점.

감사합니다