Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models

Joan Nwatu[$] **Oana Ignat**[$] **Rada Mihalcea**
University of Michigan - Ann Arbor, USA
{jnwatu, oignat, mihalcea} @umich.edu

Introduction

Digital Divide

Disparity in AI

- AI에서 소외된 소수 집단 중 저소득 가구의 데이터는 데이터 수집 및 모델 평가에서 간과되는 경우가 많음

- 기초 모델을 도출한 대부분의 연구가 서구 기술 산업에서 나온 것이기 때문에 이러한 모델은 편향된 경향성이 있음

- 기술 발전으로 인해 빈부 격차가 심화되는 상황에서, 모든 소득 수준에서 모델이 어떻게 작동하는지 명확히 이해해야 모든 경제 수준에서 공평하게 좋은 성과를 달성할 수 있음

=> Vision Language 모델에 대해 경제 수준 전반에 걸친 평가를 수행

=> Findings를 바탕으로, 모델을 '민주화'하기 위한 일련의 실행 가능한 단계를 제안함

=> 이로써 다가오는 AI 혁명의 혜택을 모두가 누릴 수 있도록 촉발시키고자 하는 것이 논문의 Motivation이자 목적

Methods

Dataset

Dollar Street Dataset

- 4개 대륙의 63개 국가로부터 수집된 38,000여 장의 일상생활 용품 이미지

- 데이터 셋에는 각각의 가구 사회경제적정보를 포함(소득, 지역 등) => 26.9$ to 19, 671.0$

- 각각의 이미지에는 text가 존재함. Text는 291개의 그림과 연관된 topic이 존재

Refrigerator

Toilet Paper

Quartile name	Income range	CLIP scores ViT-B 32	ViT-L 14	ViT-G 14
poor	26.9 - 195.0	0.233	0.259	0.321
low-mid	195.4 - 685.0	0.250	0.282	0.350
up-mid	694.0 - 1,998.0	0.257	0.295	0.363
rich	2,001.0 - 19,671.0	0.256	0.295	0.363

Table 1: Average CLIP scores per income quartile for different visual encoders.

Methods

Research Question

1. CLIP 모델의 성능이 데이터의 소득수준에 영향을 받는가?

2. 국가별, 소득별로 얼마나 많은 관련 이미지가 검색되는가?

3. 국가와 소득 수준에 따른 이미지의 주제는 국가와 소득 수준에 따라 어떤 차이가 있는가?

Experiments & Results

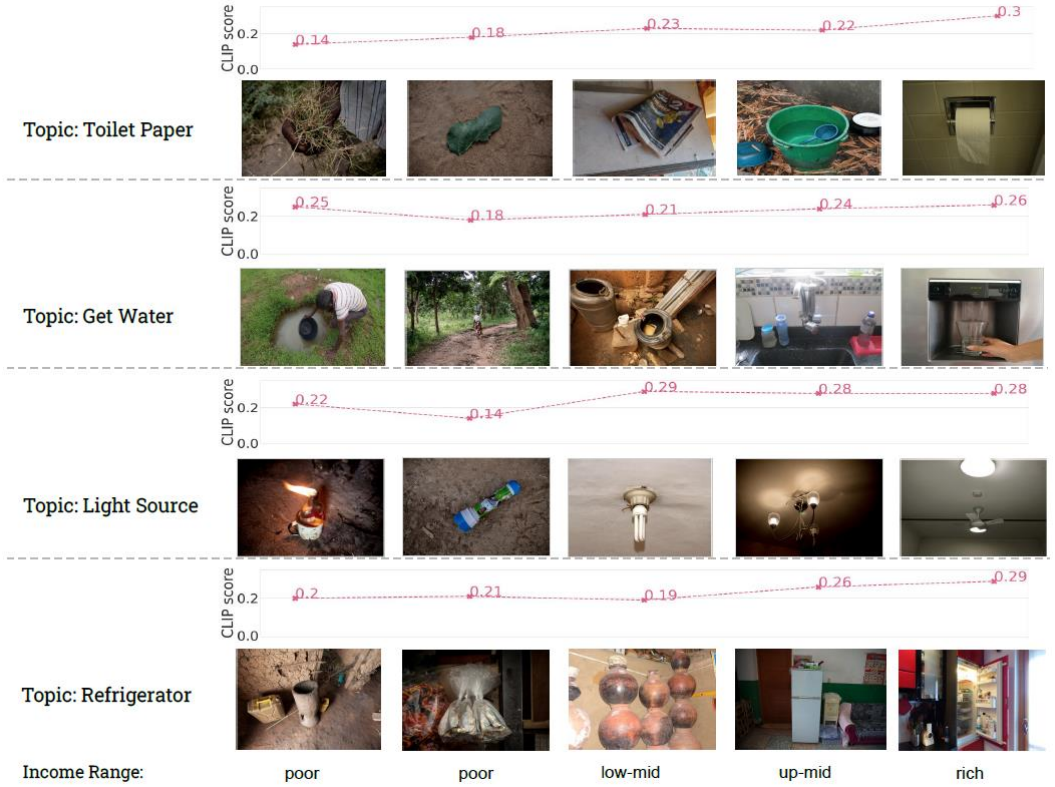
RQ1

소득 수준의 상승에 따라 CLIP score는 상승하는 경향

- CLIP은 최하 소득구간(전 세계 인구의 약 20%)의 이미지 중 약 75%에서 0.25점 미만의 성능으로 성능이 크게 저하

- 어떤 경우에 해당 구간에서도 점수가 높게 나타날까?

=> Get Water의 첫번째 이미지 처럼 Topic에 있는 단어가 이미지에 명확하게 나타나는 경우에는 높게 나타남



Experiments & Results

RQ2

모든 이미지와 주어진 Topic간의 유사도점수를 계산

- 각 Income 구간 별로 정답 이미지를 정확하게 Image Retriever를 했는지를 분석

=> 빈곤층 가구에서 수집된 이미지일 경우 검색 성능이 매우 떨어짐

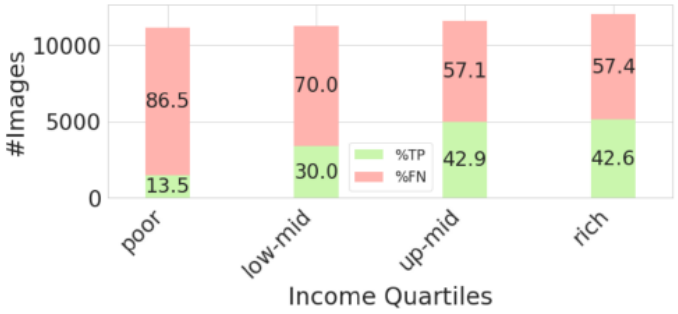
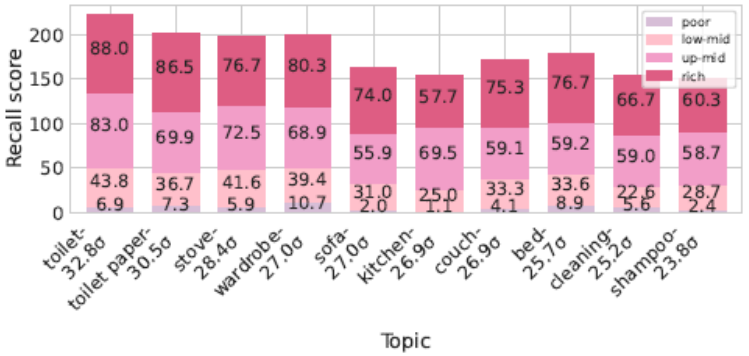


Figure 3: CLIP Recall over all images: percentage of true-positive or "recognized" images and false-negative or "forgotten" images for each income quartile. Increasingly more images are forgotten in the lower-income bins, with 86.5% forgotten images in the *poor* quartile.

어떤 Topic에서 소득수준에 따른 Recall 점수 편차가 심한가?

- "toilet", "toilet paper", "wardrobe"

=> 해당 Topic들에 대해서는 실제 사회에서 상류층과 하류층 간에 차이가 크기 때문



Experiments & Results

RQ2

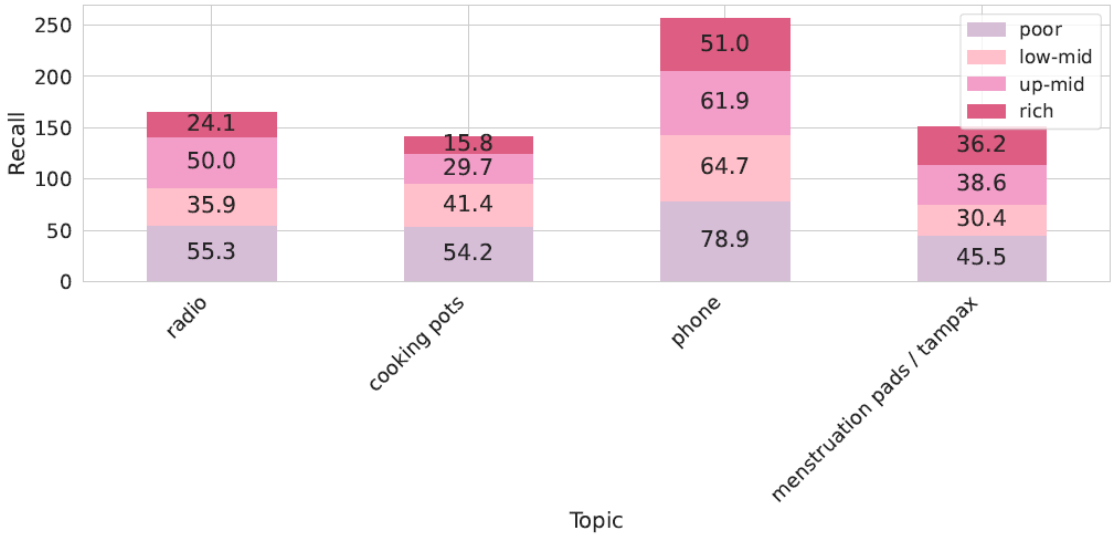


Figure 10: Topics where recall for lowest income level is higher than recall for other income levels.

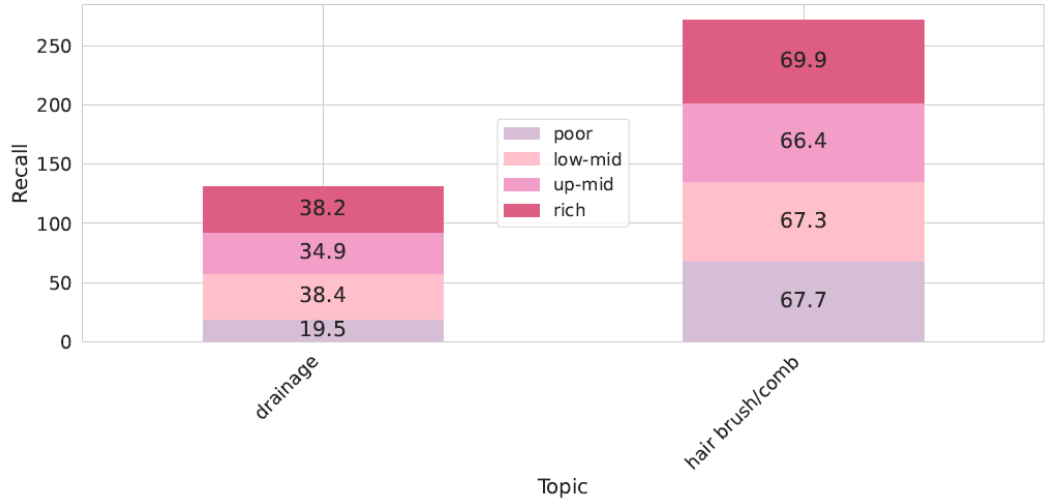


Figure 11: Topics where recall for is relatively high (hairbrush) and relatively low (drainage) across all income levels

Experiments & Results

RQ2

Disparate recall across countries

- 마찬가지로 국가별로 정답 이미지를 정확하게 Image Retriever를 했는지를 분석

=> 최악의 Recall을 기록한 대부분의 국가는 평균 소득이 낮고 아프리카에 위치

=> 저소득층, 빈곤 국가 또는 일부 지역 사회의 모습을 모델이 정상적으로 투영하고 있지 못함

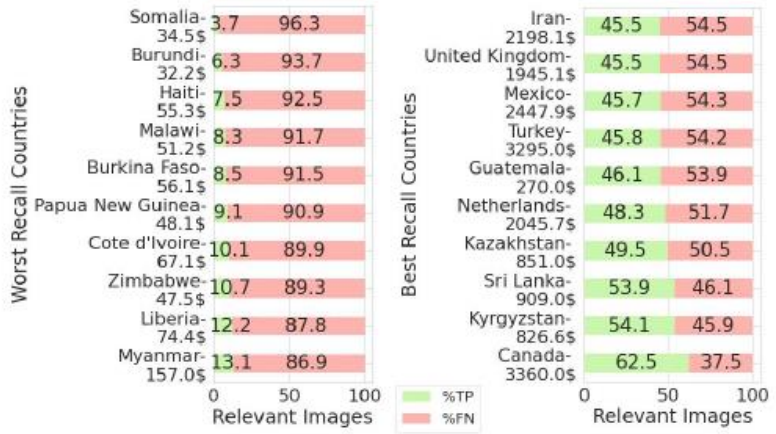


Figure 5: Countries with high and low recall scores: 7/10 countries with the worst recall have low average incomes and are from *Africa*. The countries with the best recall scores have high average incomes and are from *America, Europe, Asia*. Countries with low recall also have low income, while most countries (apart from *Guatemala*) with high recall have a high income.

Experiments & Results

RQ3

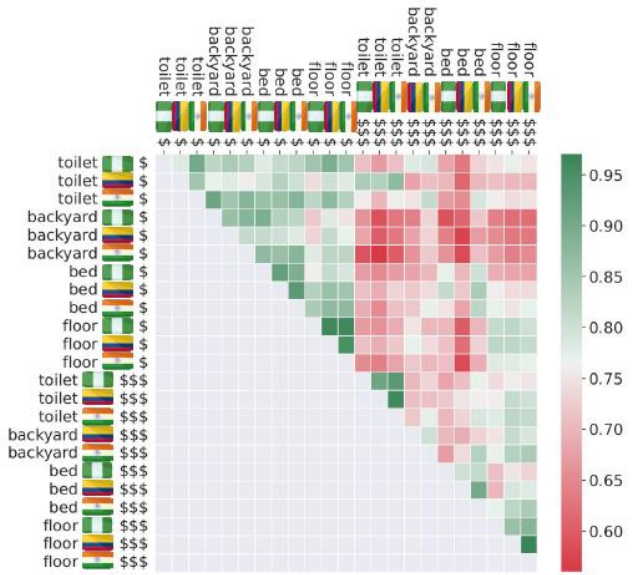
소득구간과 국가에 따라 Topic에 따른 이미지의 다양성을 분석

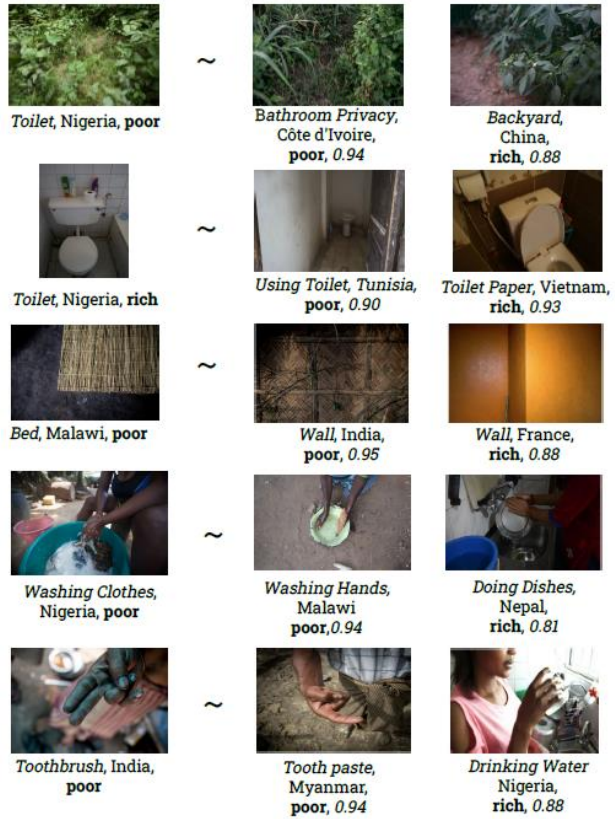
- (topic, 국가, 소득수준) 튜플을 생성하고 대응되는 이미지를 모두 임베딩

- 임베딩 간 유사도 점수가 높은 이미지가 어떤 국가, 어떤 소득 수준인지를 분석

=> 저소득층의 이미지는 같은 Topic이라면 서로 매우 유사함

=> 또한 저소득층의 이미지가 Topic에 따라 다양성이 매우 높음 (모델의 취약성에 기여)





Conclusion

Lessons Learned and Actionable Steps

1. **Invest effort to understand the extent of the digital divide in vision-language performance**
2. **Define evaluation metrics that represent everyone**
3. **Document training data**
4. **Invest in geo-diverse datasets**
5. **Annotate diversity and subjectivity in datasets**

Merging Generated and Retrieved Knowledge for Open-Domain QA

Yunxiang Zhang[*], **Muhammad Khalifa**[*], **Lajanugen Logeswaran**[†],
Moontae Lee[†‡], **Honglak Lee**[*†], **Lu Wang**[*]
University of Michigan[*], LG AI Research[†], University of Illinois at Chicago[‡]

Background

Open Domain QA

- Retrieve-then-Read Framework

⇒ Retrieving from reliable sources, enjoys the benefits of being factual

⇒ 근데...? They suffer from incomplete knowledge coverage and contain irrelevant information

- Generate-then-Read

⇒ 생성된 passage는 question과 매우 관련성이 높음

⇒ 근데...? They frequently contain factual errors due to hallucinations

“how to combine retrieved and parametric knowledge

to get the best of both worlds for open-domain QA”

Introduction

Background

Original question	*DPR-FiD* predictions	GENREAD predictions
Q: Who played lionel in as time goes by?	Geoffrey Dyson Palmer	Geoffrey Palmer

Explanation: The labeled answer is "Geoffrey Dyson Palmer", however, "Geoffrey Palmer" is also correct.
DPR retrieved documents: Geoffrey Dyson Palmer, (born 4 June 1927) is an English actor known for his roles in British television sitcoms playing Jimmy Anderson in "The Fall and Rise of Reginald Perrin", Ben Parkinson in "Butterflies" and Lionel Hardcastle in "As Time Goes By". His film appearances include "A Fish Called Wanda", "The Madness of King George", "Mrs. Brown", and "Tomorrow Never Dies".
GPT generated documents: As Time Goes By is a British sitcom that aired on BBC One from 1992 to 2005. The show starred Geoffrey Palmer and Judi Dench as Lionel and Jean Pargetter, a middle-aged couple who reunite after many years apart. Lionel was played by Palmer, who was also a writer on the show.

Methods

COMBO

Compatibility-Oriented knowledge Merging for Better Open-domain (COMBO)

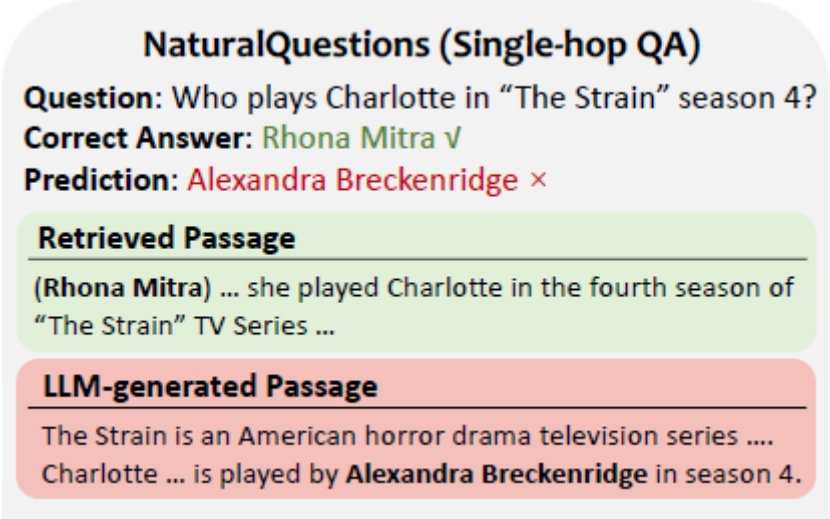
- Retrieval의 factuality + LLM의 relativeness

=> Retrieval가 가져온 문서와 LLM이 생성한 문서를 결합해
Reader 모듈에게 전달

- 결국 Reader에게 전달되는, 두 모델이 반환한 문서의 Compatibility가 중요

=> 각 프레임워크가 반환한 Passage pair에 대해
Compatibility scorer 모듈을 학습

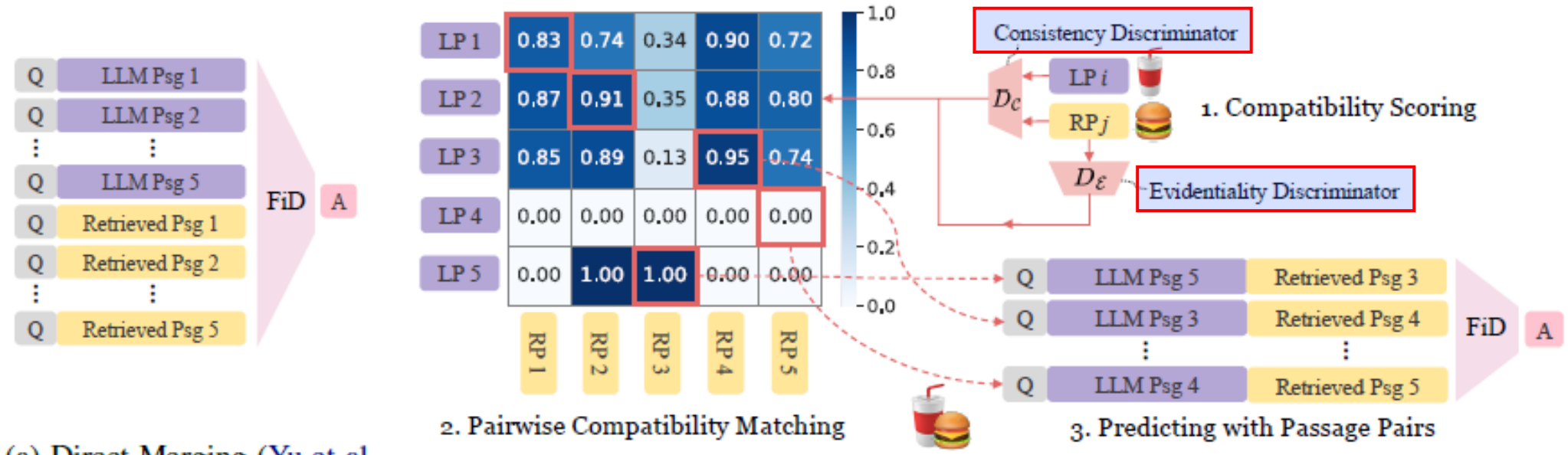
- 정한 기준을 통해 이를 학습하기위한 Silver label 데이터를 구축



Methods

COMBO

Compatibility-Oriented knowledge Merging for Better Open-domain (COMBO)



(a) Direct Merging (Yu et al., 2023)

(b) COMBO (Ours)

Methods

COMBO

Defining Compatibility

- Compatible하다

=> Passage pair가 모두 답변할 수 있는 적절한 근거를 담고 있다

$$\overbrace{P(lp_i \vDash Q, rp_j \vDash Q)}^{\text{compatibility score}} = \underbrace{P(rp_j \vDash Q)}_{\text{evidentiality score}} \cdot \underbrace{P(lp_i \vDash Q \mid rp_j \vDash Q)}_{\text{consistency score}}.$$

- Evidentiality discriminator

: Retrieved 된 Passage가 적절한 근거를 담고 있는지 판단

- Consistency discriminator

: Retrieved 된 Passage가 적절한 근거를 담고있을 때, 그에 상응하는 LLM-generated pair Passage 또한 적절한 근거를 담고 있는지 판단

=> rp_j가 evidential하다면, rp_j + Q + lp_i를 input으로 넣고 둘의 부합 여부를 분류

결국 반환된 passage pair가 답변하기 적절한 정보를 담고 있는지 구분하는 것을 학습!

Methods

COMBO

Mining Silver Labels

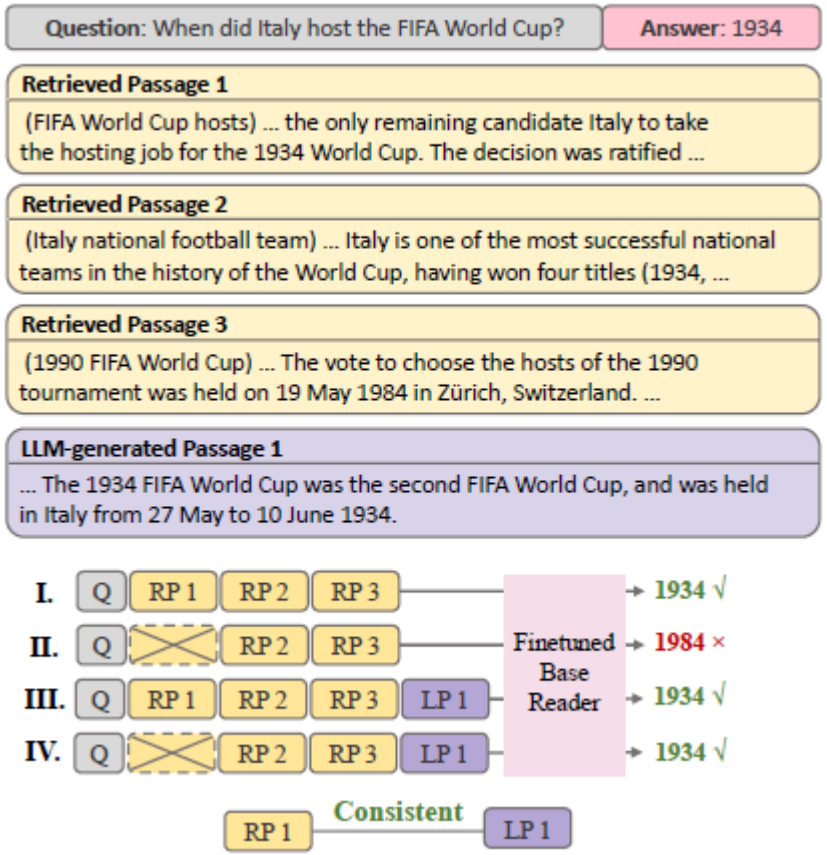
- I: 모든 RP를 포함

- II: target RP를 제외한 모든 RP를 포함

- III: 모든 RP와 target LP를 포함

- IV: target RP를 제외한 모든 RP와 target LP를 포함

• RP: Retriever가 뽑은 passage
• LP: LLM이 생성한 passage



Methods

COMBO

Mining Silver Labels

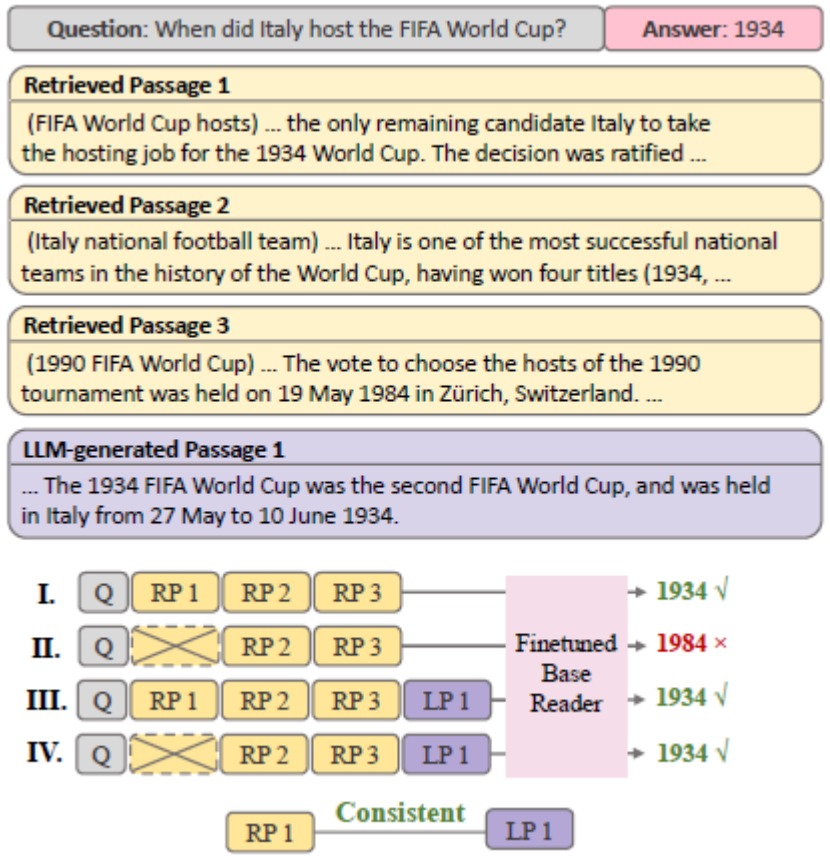
- 어떻게 하면 Evidentiality, Consistency discriminator를 학습시킬 gold를 만들까?

- Evidential

: QA model에 input되는 passage 중 해당 passage를 제외했을 때 correct -> incorrect (I-> II)

- Consistent

: Retrieved 된 Passage가 제외되었을 때 정답이 변하지 않고 (I -> II), LLM-generated Passage가 제외되었을 때 정답이 틀려지는 경우 (IV -> II)



Experiments

Open-domain QA

Comparison with Baselines

- 모두 각각 10개의 retrieved passage, LLM-generated passage를 사용

- 각각의 경우보다 성능이 큰 폭으로 증가

- 랜덤 또는 직접적으로 매칭하는 것보다 효과적

- But, HotPotQA와 같은 multi-hop setting에서는 미미함

=> 증거 홉과 compatibility간의 연결성을 매칭하지 않았기 때문

Methods	NQ test	TQA test	WebQ test	HQA all Q dev	HQA bridge Q dev
Single Knowledge Source					
Retrieved Psg. Only	46.7	61.9	48.1	59.9	55.4
LLM Psg. Only	40.3	67.8	51.5	42.6	35.9
Two Knowledge Sources					
Direct Merging	52.7	74.2	51.1	**61.6**	57.8
Random Matching	53.3	74.2	51.6	61.5	57.7
COMBO (ours)	**54.2**	**74.6**	**53.0**	**61.6**	**58.0**

Table 1: Exact Match (EM) results on NaturalQuestions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestion (Berant et al., 2013), and HotpotQA (Yang et al., 2018). For experiments under *Two Knowledge Sources*, we conduct three runs with different random seeds and report the average.

Experiments

Open-domain QA

Ablation Study

Method	NQ dev	TQA dev
COMBO (ours)	**52.3**	**73.9**
w/o evidentiality discriminator	51.8	73.5
w/o pairwise input	51.5	73.4
w/o sorting pairs	51.7	73.7
w/o fixed (lp_i, rp_j) order	50.8	73.3
w/o evidentiality-cutoff	51.7	73.6
w/o optimal matching	51.6	73.6

Table 2: Ablation study results on NaturalQuestions (NQ) and TriviaQA (TQA).

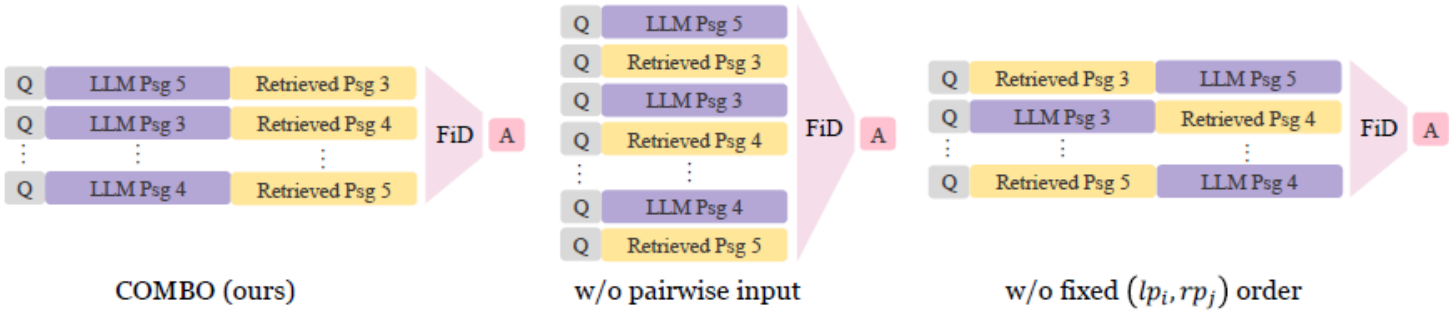


Figure 4: Illustrations of the input formats for two of the ablation experiments in Table 2.

Experiments

Open-domain QA

Impact of Conflicting Contexts on the Reader Model

- LLM과 Retrieval 사이의 충돌이 있을 때, Reader가 정답을 얼마나 잘 찾아낼 수 있는가?

- Conflicting rate

$$\text{conflicting\_rate} = \frac{N_A \cdot (M - M_A)}{N \cdot M}.$$

N_A: # of Gold answer를 포함한 retrieved Passage

$M\text{-}M_A$: # of gold answer를 포함하지 않은 LLM Passage

=> Conflict rate이 높아도, Direct Merging보다 일관된 성능 향상이 관찰

Conflicting Rate	Subset%	Retrieved Psg. Only	Direct Merging	COMBO
0 – 0.1	56.2%	41.6	47.7	**48.5** (+0.8)
0.1 – 0.2	22.7%	45.1	53.1	**53.3** (+0.2)
0.2 – 0.3	15.5%	52.5	59.0	**59.9** (+0.9)
0.3 – 0.4	1.8%	62.5	65.0	**67.5** (+2.5)
0.4 – 0.5	2.2%	57.4	64.0	**66.0** (+2.0)
0.5 – 1.0	1.6%	61.8	56.6	**61.0** (+4.4)

Table 3: The performance of our method compared to others on NaturalQuestions dev set w.r.t. conflicting rate (percentage of conflicting passage pairs, Equation 3). Larger improvements of COMBO over Direct Merging are shaded with darker orange. Overall, COMBO's improvement over Direct Merging is greater when the conflicting rate is higher, suggesting the robustness of our method to knowledge conflicts.

Experiments

Open-domain QA

Case Study

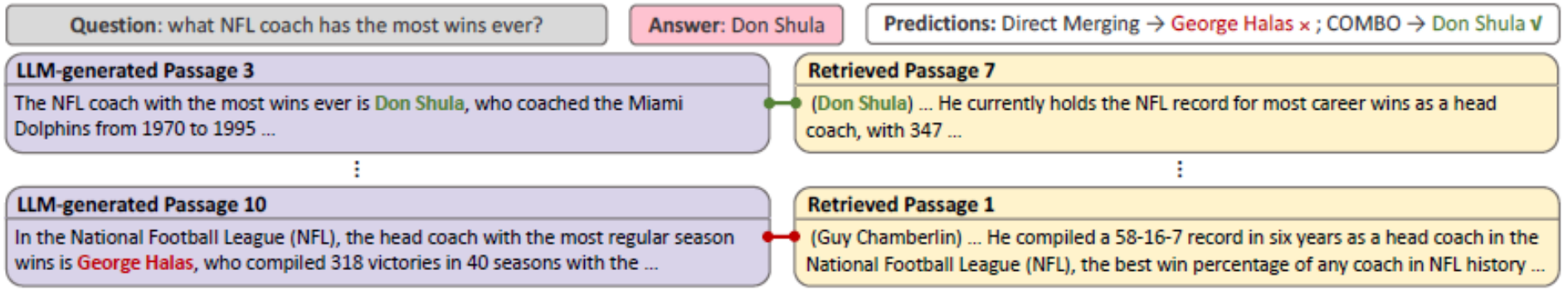


Figure 6: An example of a QA pair and the passage matching results by COMBO. Passage pairs are sorted by their compatibility scores. It shows how COMBO rectifies the prediction of the baseline method under knowledge conflicts by prioritizing compatible pairs (green connecting line) over incompatible pairs (red connecting line).

- LLM-generated passage는 hallucination으로 George Halas를 답변으로 생성

- COMBO의 경우 Direct Merging에서 틀렸던 정답을, compatible pair를 우선순위화 시킴으로써 올바른 정답으로 교정함

Experiments

Open-domain QA

Human Evaluation

- Discriminator가 passage의 compatibility를 reader에게 적절히 전달하고있는가

<=> Silver Labeling이 정상적인가

- compatible pair가 실제로 compatible한지, evidential pair로 판별된 pair가 정답 근거를 포함하고 있는지를 human annotating

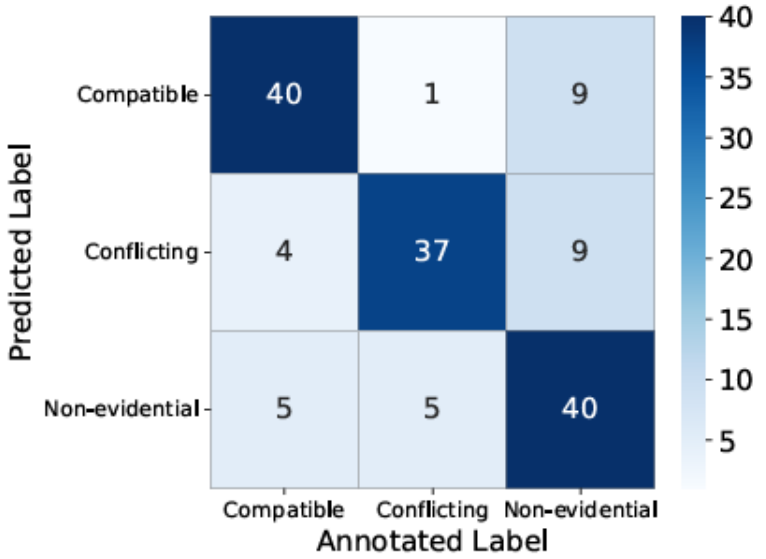


Figure 7: Confusion matrix of the human annotations vs. predicted labels by our discriminators on 150 random samples from NaturalQuestions dev set. The overall accuracy is 78%.

Conclusion

- **Retrieved passage는 이런 문제, LLM-generated passage는 이런 문제가 있다**

: Knowledge conflict & Compatibility

- **실험이 많은 편은 아님. 자신들의 논리내에서 부합하는 정말 필요한 실험들만 보여줌**

: Knowledge conflict 이만큼 발생한다 / Compatibility 유효하다 / 심지어 Conflicting 상황에서도 좋다

- **Appendix**

: Appendix에 Retrieved-passage, LLM-generated passage에 대한 Preliminary Study를 포함

- **Limitation이 자세한편**

: 자신들의 과업범위 명확히. 더 intensive한 확장 태스크들에 대해서는 it would be interesting. 예상 가능한 공격에 대해 사전 방어

Thank you

Q&A