KAIST AI
Kim Jaechul Graduate School

Knowledge Corpus Error in Question Answering

Yejoon Lee, Philhoon Oh, James Thorne

Findings of EMNLP 2023

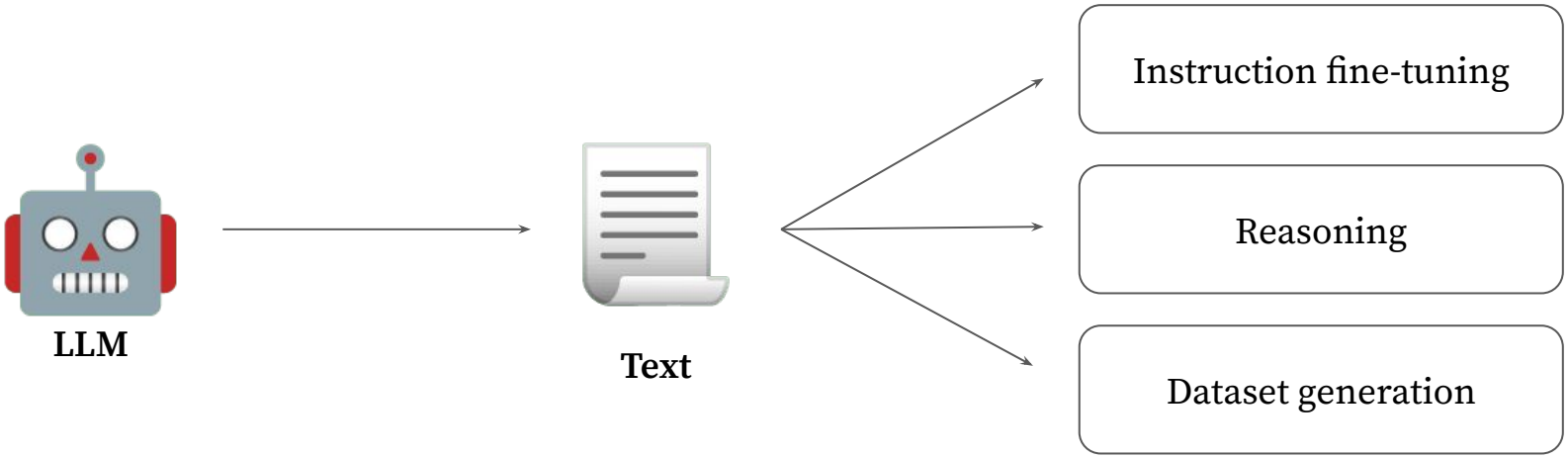
leeyejoon@snu.ac.kr

https://xfact.net

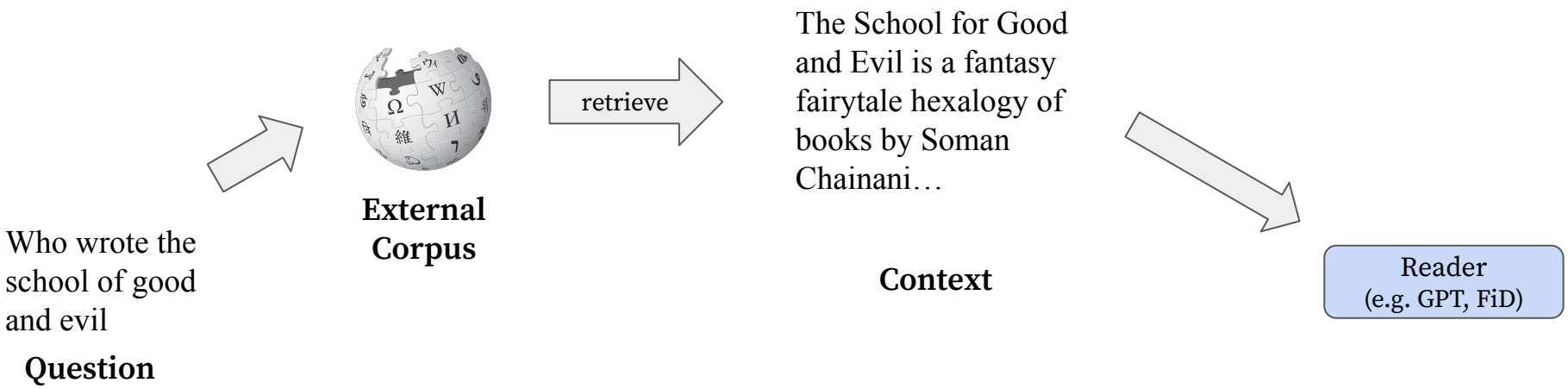
Motivation

LLMs generate fluent and informative texts

Wang et al. ACL 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions
Zelikman et al. NeurIPS 2022. STaR: Self-Taught Reasoner Bootstrapping Reasoning With Reasoning
Liu et al. EMNLP Finding 2022. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation

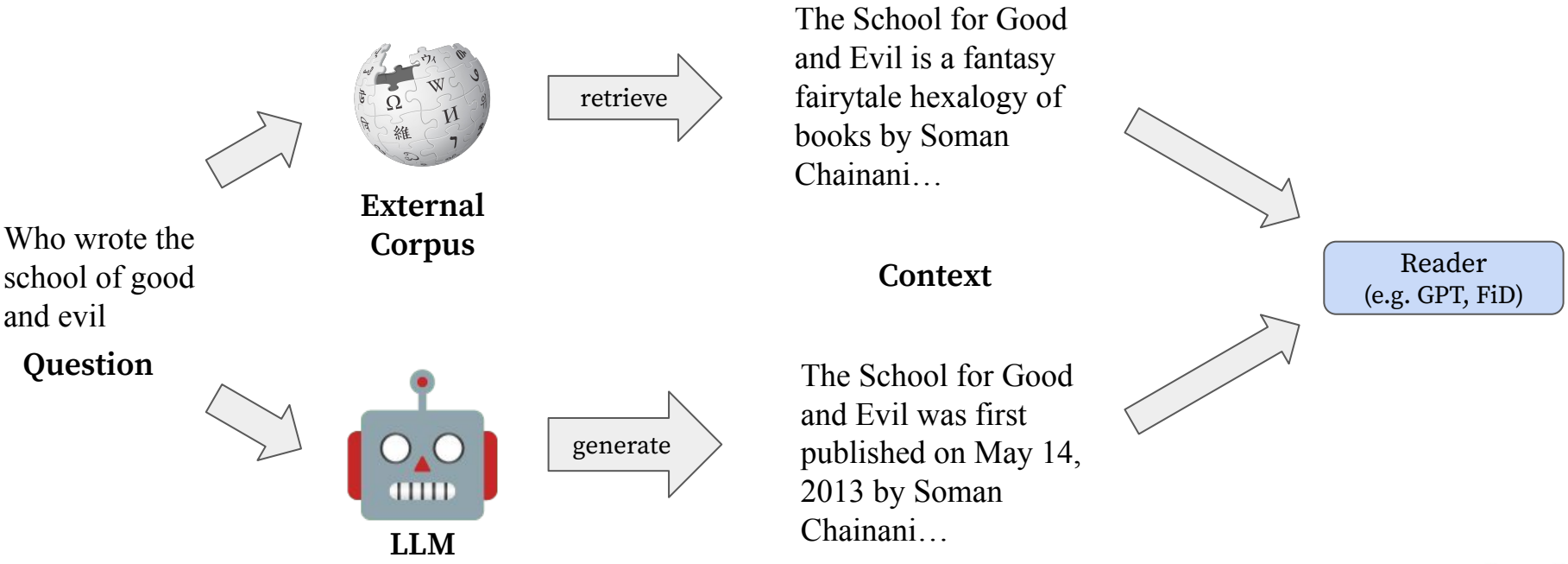


Retrieve-then-read pipeline in QA

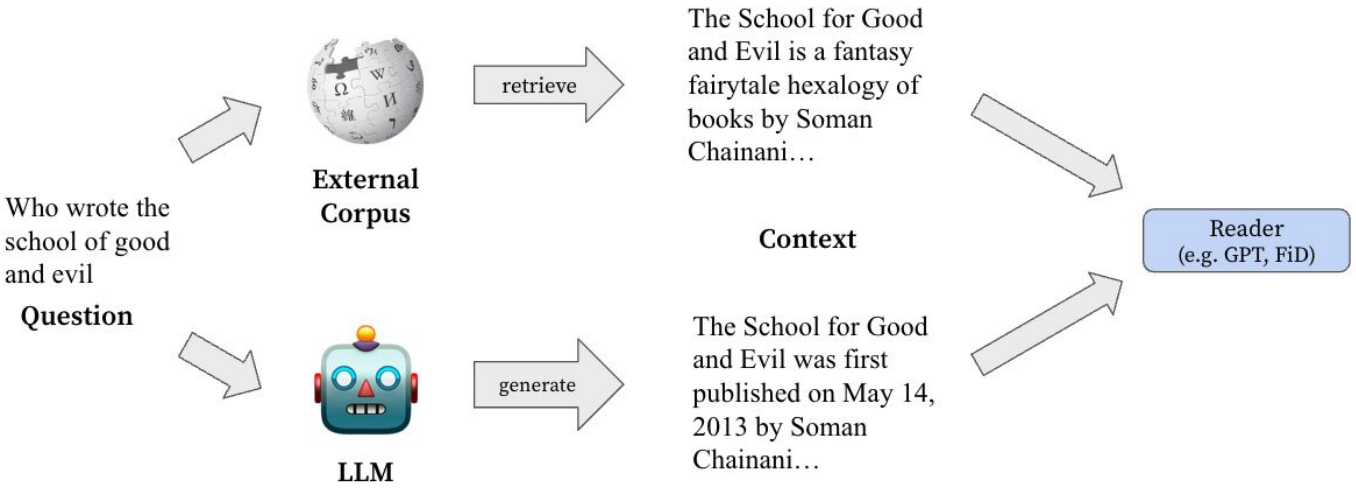


LLMs can generate context

Yu et al. ICLR 2023. Generate rather than Retrieve: Large Language Models are Strong Context Generators
Sun et al. ICLR 2023. Recitation-Augmented Language Models



Motivation



- Not well understood why generated passages could be more effective
- Lack robust links to prior research in QA

Formulation

Typical formulation of QA

Guu et al. ICML 2020. Retrieval Augmented Language Model Pre-Training
Lewis et al. NeurIPS 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
Singh et al. NeurIPS 2021. End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering

$$p(a|q)$$

Learn a distribution

$$\hat{a} = \arg\max_{a \in V^*} p(a|q)$$

Decode a string *a* that acts as an answer

Typical formulation of QA

Guu et al. ICML 2020. Retrieval Augmented Language Model Pre-Training
Lewis et al. NeurIPS 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.
Singh et al. NeurIPS 2021. End-to-End Training of Multi-Document Reader and Retriever for Open-Domain Question Answering

$$p(a|q)$$

Learn a distribution

$$\hat{a} = \arg\max_{a \in V^*} p(a|q)$$

Decode a string *a* that acts as an answer

$$p(a|q) \approx \sum_{c \in Topk(\mathcal{Z}, q)} p(a|q, c)p(c|q)$$

Marginalize over contexts in the knowledge corpus $\mathcal{Z}$

The gap

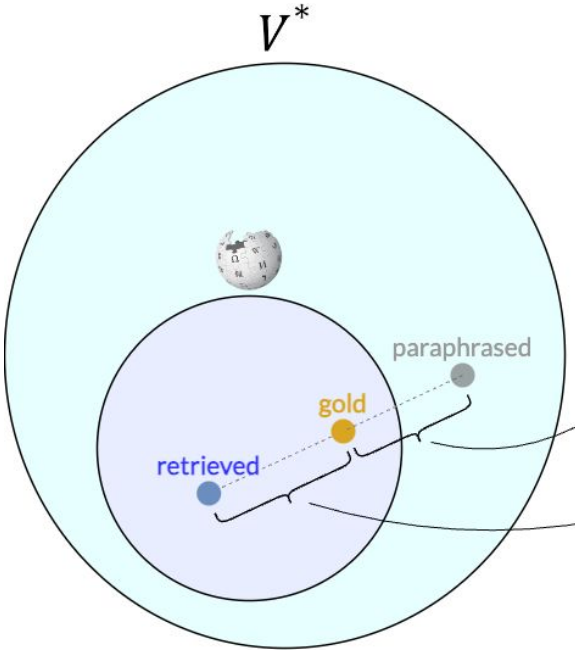
$$p(a|q) = \sum_{c \in V^*} p(a|q,c)p(c|q)$$

$$\approx \sum_{c \in \mathcal{Z}} p(a|q,c)p(c|q)$$

$$\approx \sum_{c \in Topk(\mathcal{Z},q)} p(a|q,c)p(c|q)$$

Helpful context *c* may exist **outside** of knowledge corpus!

Knowledge Corpus Error



knowledge corpus error

retrieval error

$$p(a|q) = \sum_{c \in V^*} p(a|q,c)p(c|q)$$

$$\approx \sum_{c \in \mathcal{Z}} p(a|q,c)p(c|q)$$

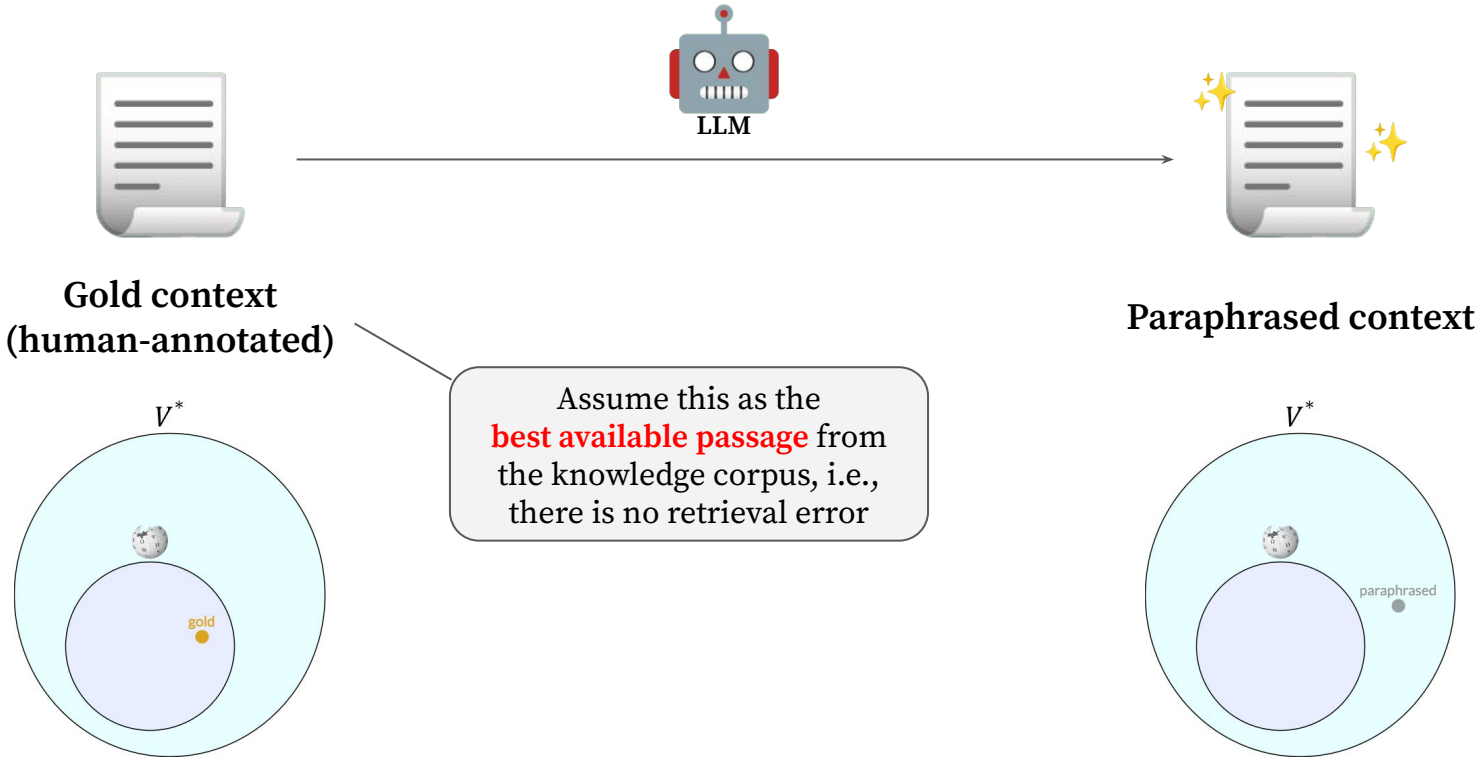
$$\approx \sum_{c \in Topk(\mathcal{Z},q)} p(a|q,c)p(c|q)$$

Experiments

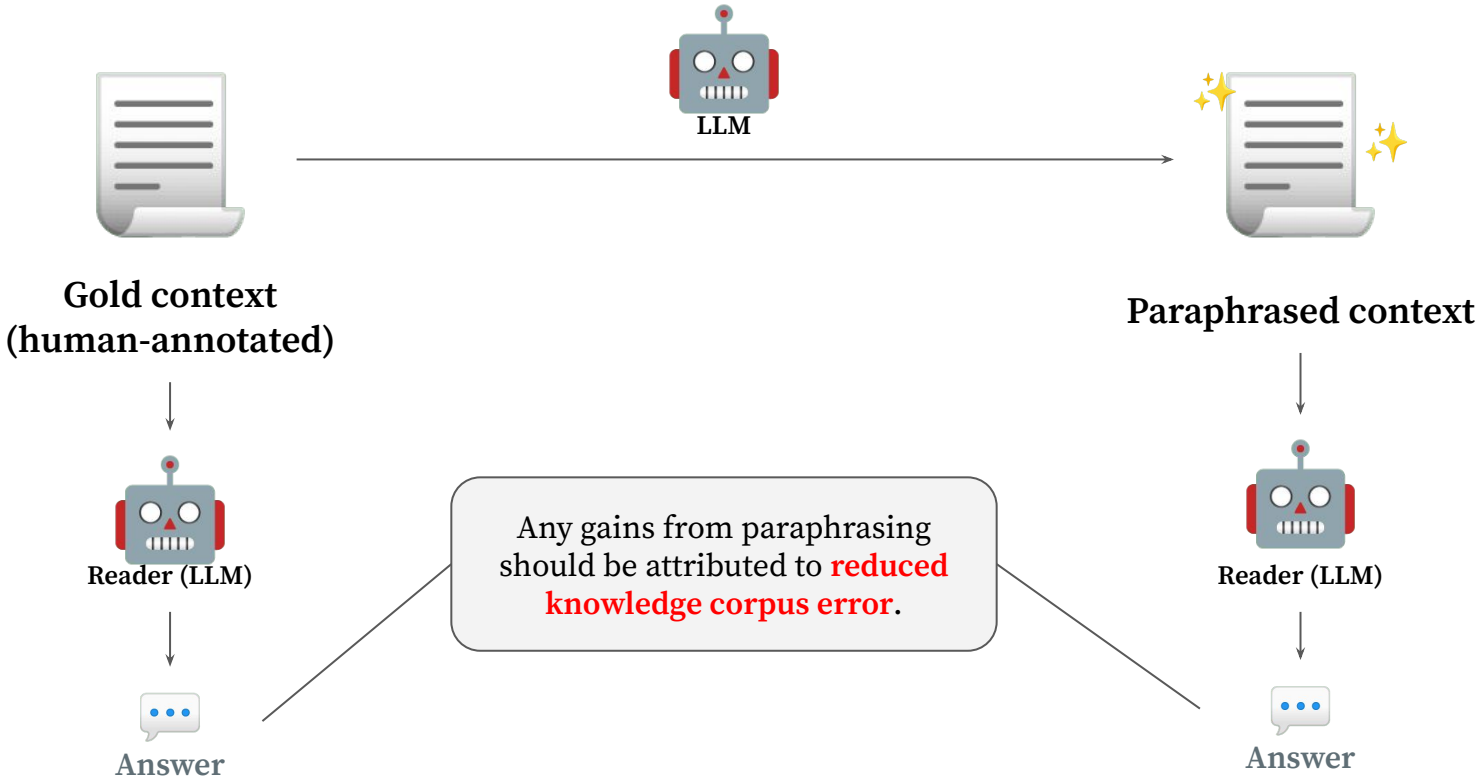
Goal of experiments

Question: Can we empirically observe knowledge corpus error?

Experimental setup



Experimental setup



Results

Benchmarks	Reader: GPT			Reader: Claude			Average gap between gold and paraphrased
	Gold	GPT	Claude	Gold	GPT	Claude	
NQ exact match (%)	40.9	39.9	44.3	18.3	21.3	35.5	3.125
HotPotQA exact match (%)	36.3	38.6	43.4	47.6	50.9	54.2	4.825
StrategyQA accuracy (%)	54.6	56.4	70.5	68.9	75.5	76.5	7.975
QASC accuracy (%)	95.7	92.4	91.1	86.3	75.7	76.9	- 6.975

Table 1: Performance of each reader when given original gold context ("Gold"), paraphrased context with GPT ("GPT"), and paraphrased context with Claude ("Claude"). Red indicates an increase in performance after paraphrasing, implying knowledge corpus error has been observed. Blue indicates a decrease in performance after paraphrasing, implying knowledge corpus error has not been observed.

- Gain in performance across 3 benchmarks
- Degradation in QASC is excusable (check the paper for details)

Thank You

xfact