NAACL 2024, Findings

Discern and Answer: Mitigating the Influence of Noise on Retrieval-Augmented Models with Discriminators

Giwon Hong[1*], Jeonghwan Kim[2*], Junmo Kang[3], Sung-Hyon Myaeng[4], Joyce Jiyoung Whang[4]

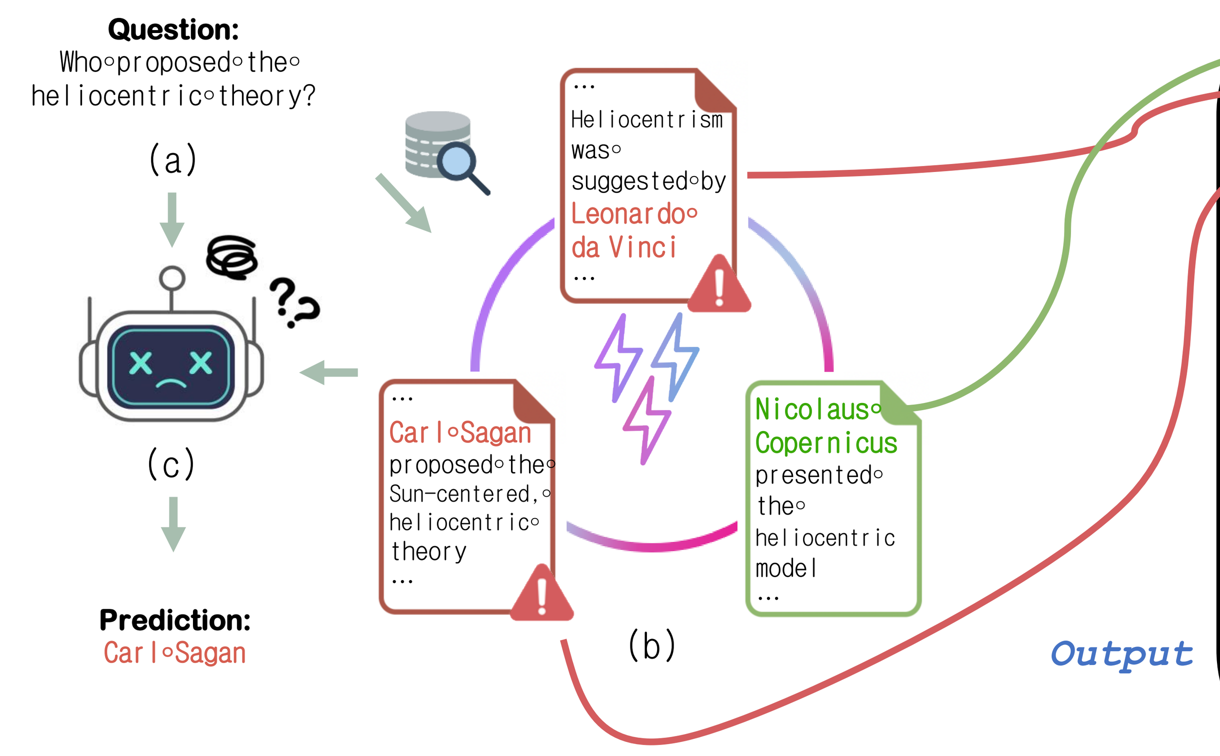
[1]University of Edinburgh [2]University of Illinois Urbana-Champaign [3]Georgia Institute of Technology [4]KAIST

* Work was done while working at KAIST

Paper Link

Motivation

- Misinformation and its impact on the Web are ever-increasing (Vicario et al., 2016)



- We focus on handling misinformation in a set of retrieved documents in **open-domain question answering (ODQA) setting**

Preliminarily Study

- Misinformation can be detrimental, especially for LLMs, which are challenging to fine-tune

```
Passage 1: <TITLE> context: <CONTEXT>
Passage 2: <TITLE> context: <CONTEXT>
…
Passage N: <TITLE> context: <CONTEXT>

Some passages may have been perturbed with
wrong information. Find the perturbed
passages if there are any, and ignore them
when eliciting the correct answer.

Question: Who proposed the heliocentric
theory?
Perturbed: Passage 1, … N-1 are perturbed.
Deriving the answer based on Passage 2, N
Answer: Carl Sagan
```

- With in-context learning, we can simply detect misinformation before generating answers

- However, **LLMs** exhibit limited ability to classify misinformation

GPT-3.5

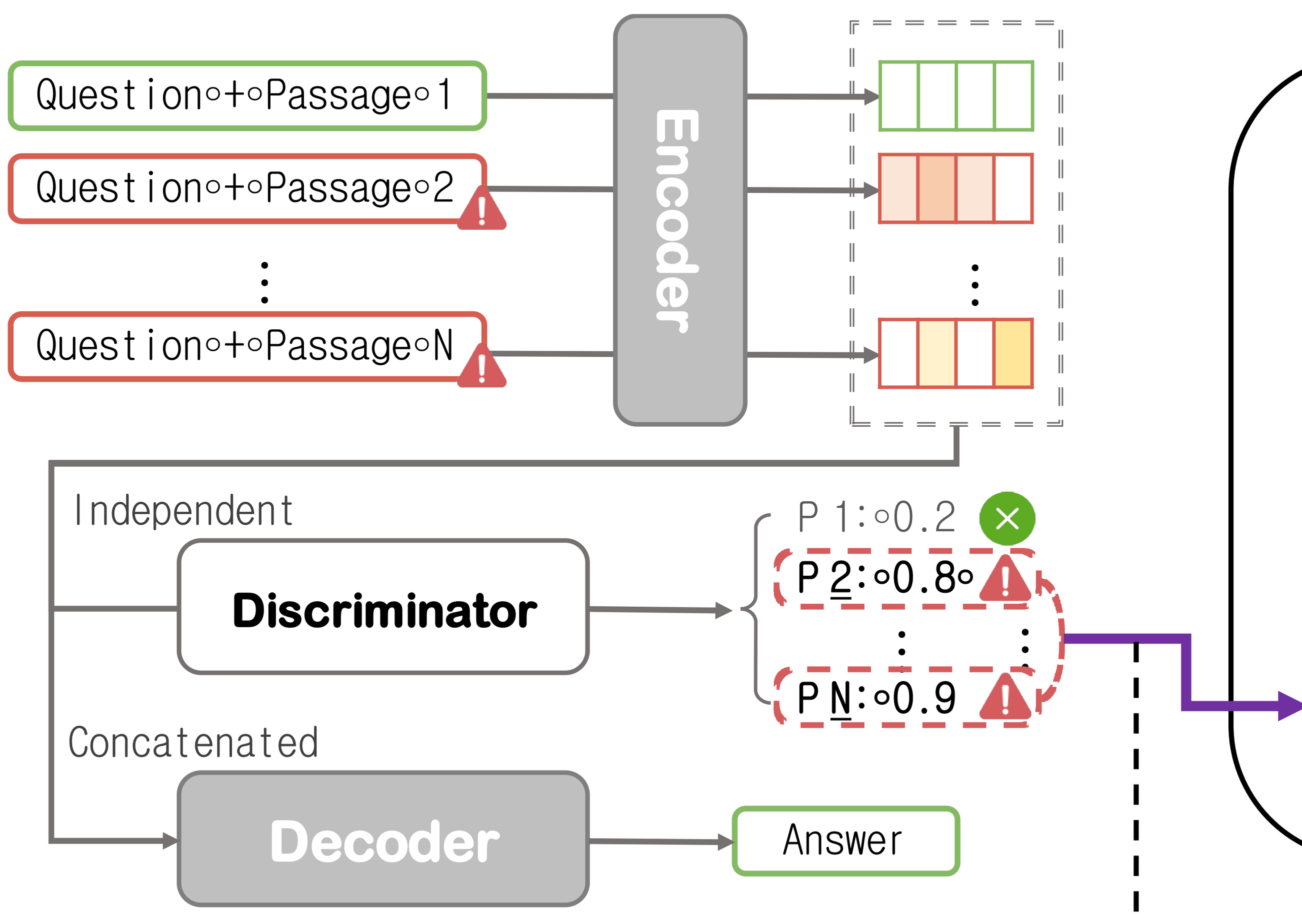
Mis.%	Prec.	Rec.	F1
15%	20.14	49.11	28.57
25%	30.29	48.59	37.32
35%	42.03	49.14	45.31

- Meanwhile, smaller **fine-tuned models** show better classification abilities

Fine-tuned T5-base

Mis. %	Prec.	Rec.	F1
15%	93.60	61.26	74.05
25%	98.51	63.78	77.43
35%	96.28	68.65	80.15

Proposed Method




```
Passage 1: <TITLE> context: <CONTEXT>
Passage 2: <TITLE> context: <CONTEXT>
…
Passage N: <TITLE> context: <CONTEXT>

Some passages may have been perturbed with
wrong information. Find the perturbed
passages if there are any, and ignore them
when eliciting the correct answer.

Question: Who proposed the heliocentric
theory?
Perturbed: Passage 2, … N are perturbed.
Deriving the answer based on Passage 1, N-1
Answer: Nicolaus Copernicus
```

- A fine-tuned model (FiD; Izacard et al., 2021) specialized for misinformation

Inject classification results of the fine-tuned model into LLM's prompts

Settings

- **Task: Open-Domain QA**
  - Natural Questions(NQ) (Kwiatkowski et al., 2019)
  - TriviaQA (Joshi et al., 2017) (omitted)
- **Entity Perturbation Method**
  - Longpre et al. (2021)
- **LLM-generated perturbation**
  - **MacNoise**
- **Models**
  - Fine-tuned Model: Fusion-in-Decoder (FiD)
  - LLM: GPT-3.5

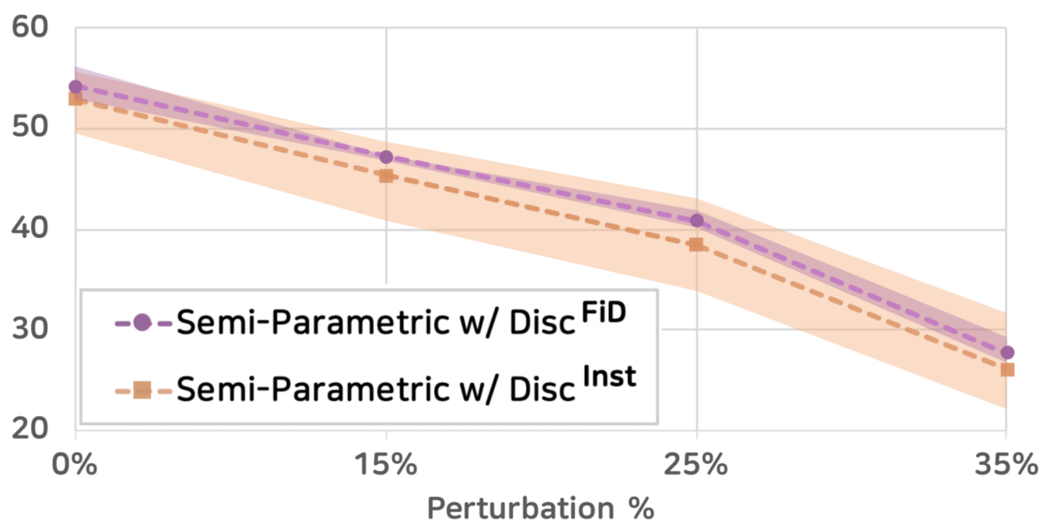
Results

- $Disc^{Inst}$: Instruction-based classification (preliminarily study)
- $Disc^{FiD}$: Fine-tune model's classification (proposed method)

Method	Perturbation % (Dev / Test)				
	0%	15%	25%	35%	Avg.
Parametric (w/o Retrieval)	32.0 / 36.8				32.0 / 36.8
Semi-Parametric (w/ Retrieval)	50.4 / 53.2	40.2 / 45.0	31.3 / 37.8	22.7 / 24.2	36.2 / 40.1
Semi-Parametric w/ $Disc^{Inst}$	48.8 / 54.2	37.9 / 45.6	28.9 / 38.4	21.5 / 26.8	34.3 / 41.3
Semi-Parametric w/ $Disc^{FiD}$	**51.2 / 56.3**	**42.2 / 49.2**	**34.0 / 41.6**	**27.3 / 28.6**	**38.7 / 43.9**
Δ Absolute Gain	+0.8 / +3.1	+2.0 / +4.2	+2.7 / +3.8	+4.6 / +4.4	+2.5 / +3.8

- Due to LLM's inferior misinformation detection ability, $Disc^{Inst}$ does not show performance improvement
- By utilizing predictions from the specialized fine-tuned models, $Disc^{FiD}$ **shows consistent performance improvement**
- Nevertheless, if the retrieved documents are severely contaminated, it is better to rely solely on parametric knowledge

Enhanced In-Context Learning Stability



- Utilizing the fine-tuning model's predictions significantly **reduced variance across different examples of in-context learning**

MacNoise: Machine-Generated ODQA Benchmark

Original Document from Natural Questions (NQ)

… the company is now the largest American retailer of women's lingerie. Victoria's Secret was founded by **Roy Raymond**, and his wife **Gaye Raymond** …

MacNoise

Context: Victoria's Secret is an American designer, manufacturer, and marketer of women's lingerie, womenswear, and beauty products. The company was founded in 1977 by **John Thompson** and his wife, **Gaye Thompson**, in San Francisco, California …

Conclusion

- In-context learned LLMs are brittle to the presence of misleading information
- Our approach significantly enhances the LMs' ability to handle conflicts
- We present **MacNoise**, a novel knowledge conflict ODQA benchmark
- Combining the fine-tuned model's output with in-context learning, **creating a new avenue for future work to harness the advantages of both learning paradigms**