Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models against Counterfactual Noise

NAACL 2024, Findings

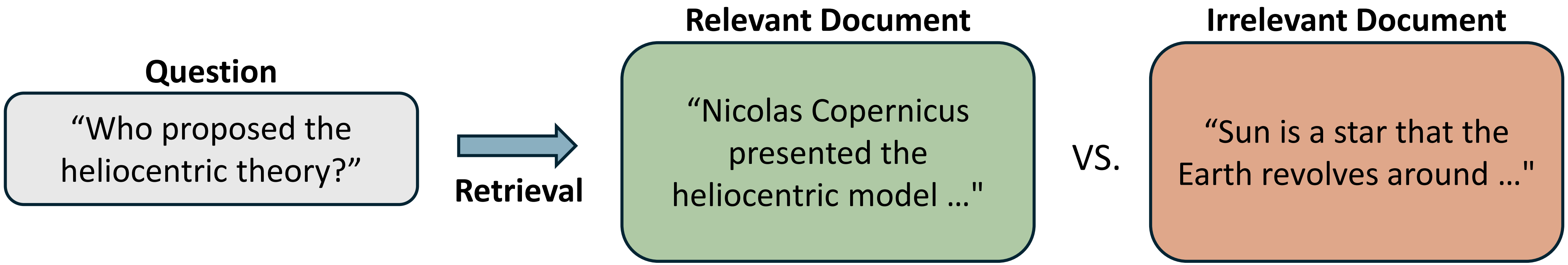
Giwon Hong*, Jeonghwan Kim*, Junmo Kang*,
Sung-Hyon Myaeng, Joyce Jiyoung Whang

THE UNIVERSITY of EDINBURGH
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN
Georgia Institute of Technology
KAIST

Motivation

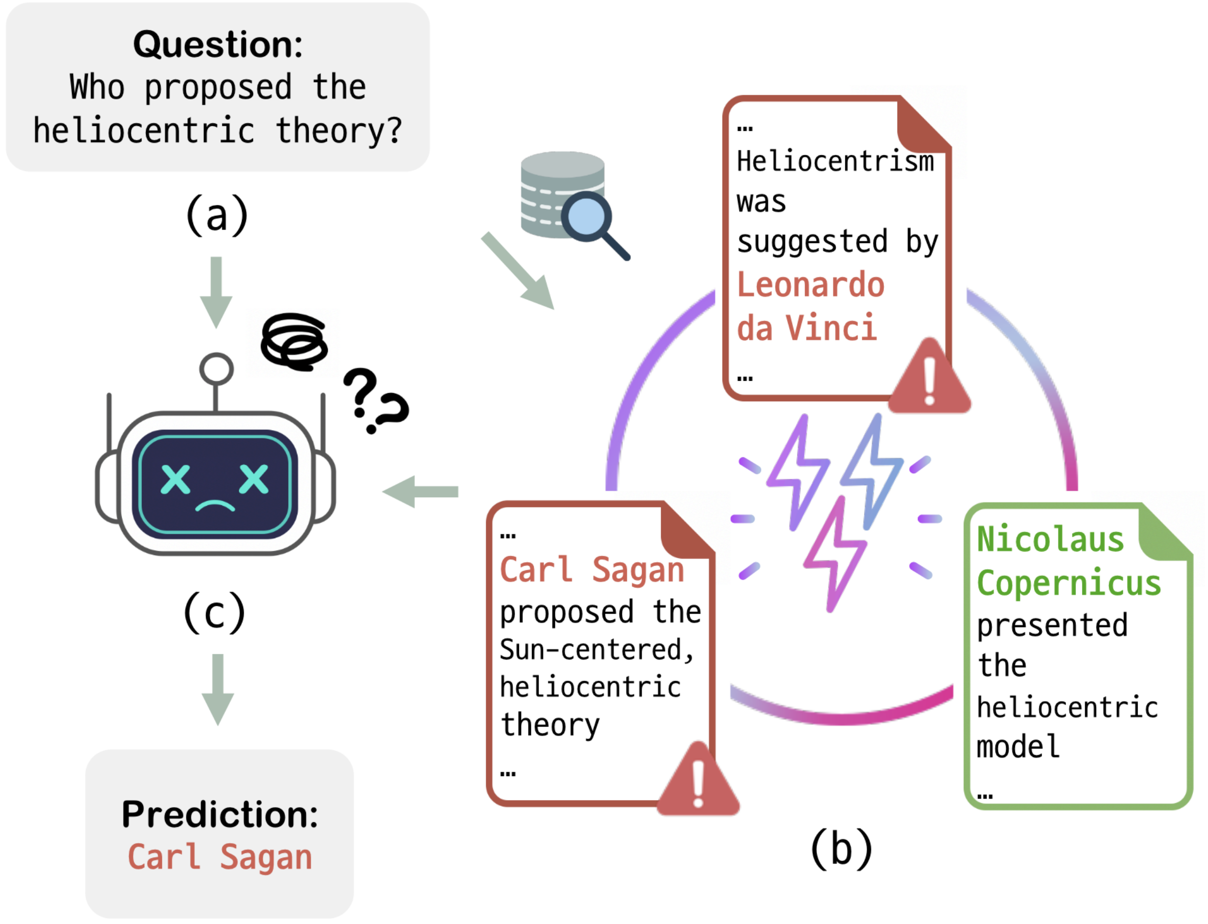
Retrieval-Augmented Language Models (RALMs) often assume a naïve dichotomy among retrieved documents

Relevance vs. Irrelevance



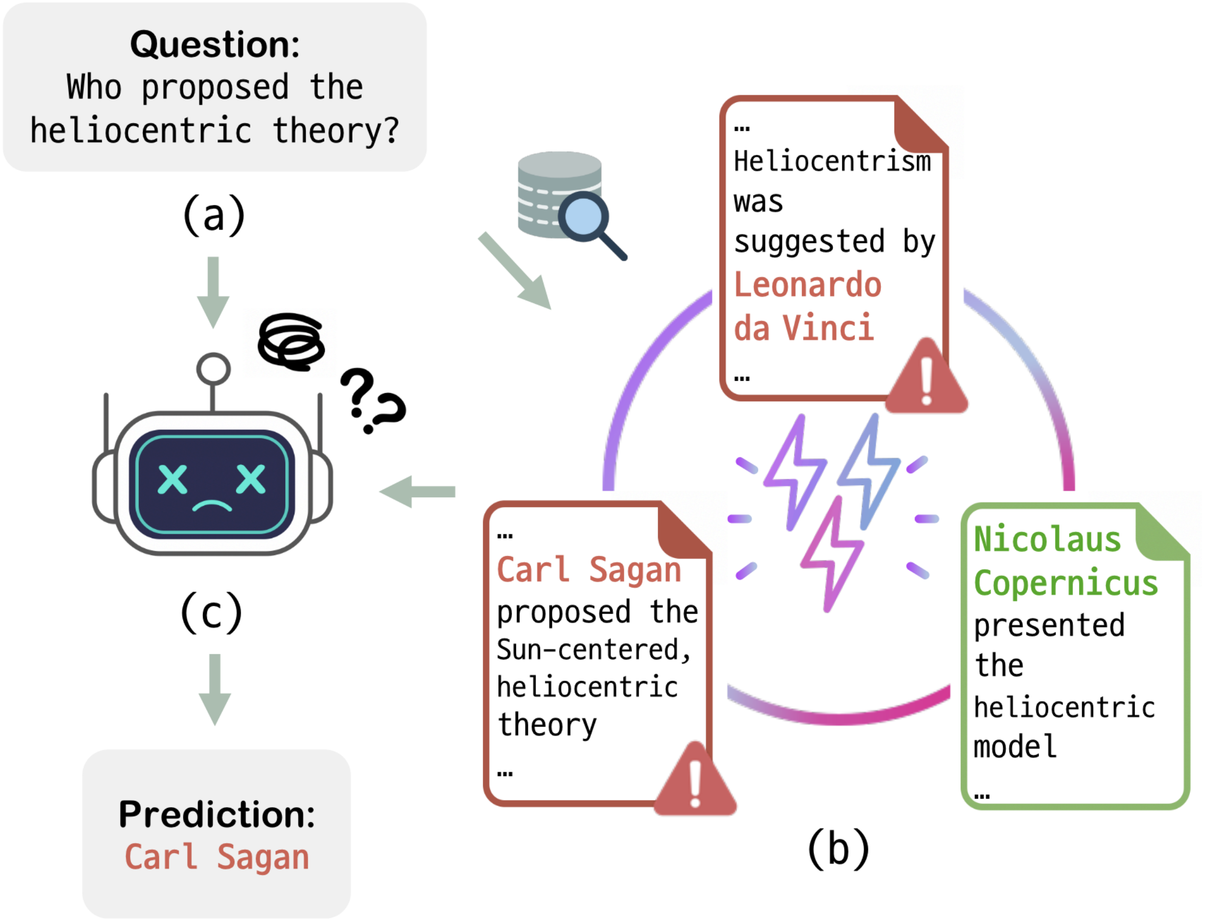
Overview

Our work studies a more challenging scenario in Open-Domain Question Answering (ODQA), wherein the retrieved relevant documents contain **counterfactual noise**

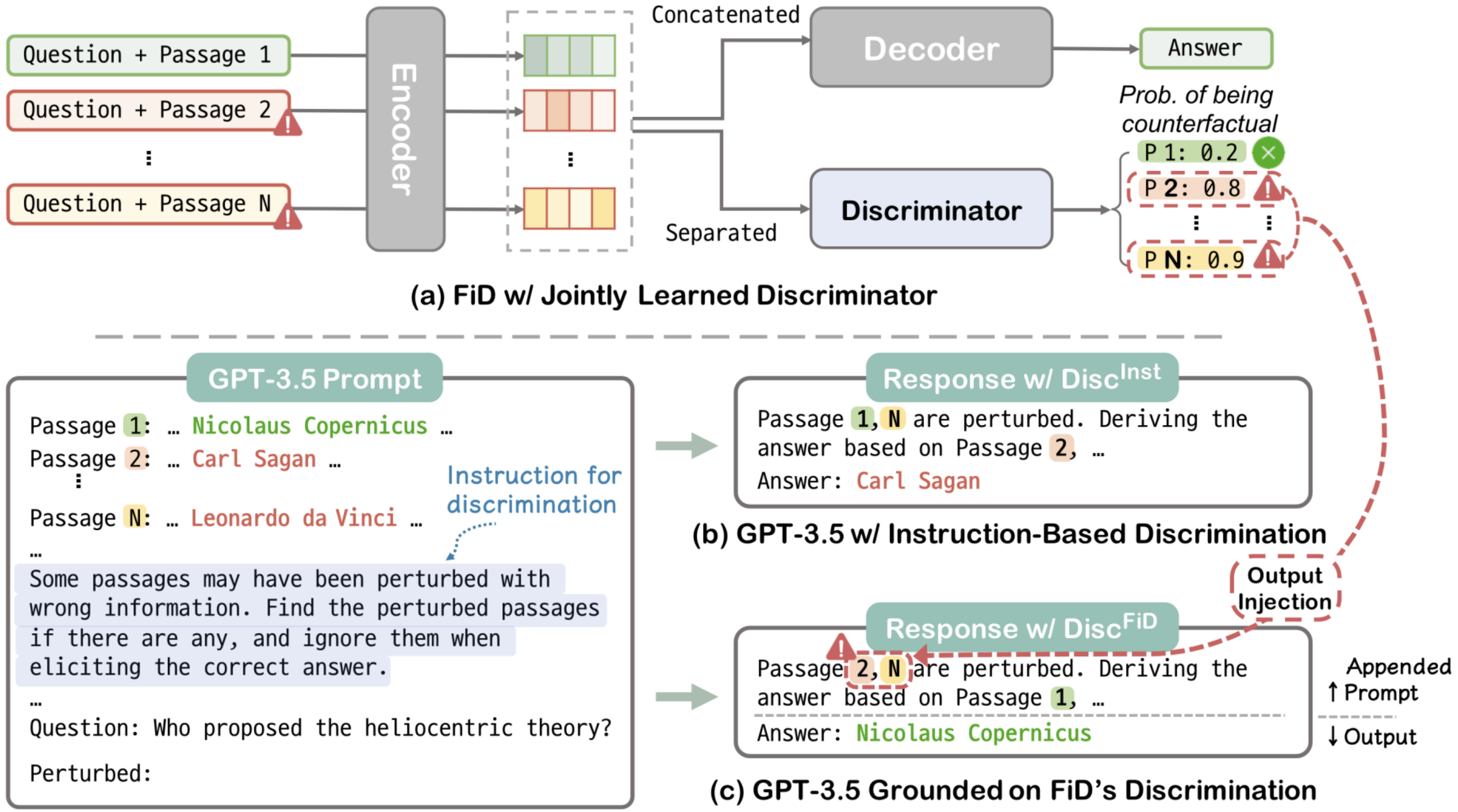


Overview

We investigate the **robustness of RALMs** given a retrieved set of **counterfactual** and **gold** documents in ODQA **(knowledge conflict)**



Overview



We propose a simple yet effective approach to enhance the **discriminative capabilities** of RALMs such as FiD[1] and GPT-3.5[2]

[1] Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering, Izacard et al., EACL 2021
[2] Language Models are Few-Shot Learners, Brown et al., NeurIPS 2020

Overview

Original Document from Natural Questions (NQ)[1]

… 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** …

Entity-Centric Perturbation [2]

… the company is now the largest American retailer of women's lingerie. Victoria's Secret was founded by **Patrick Denham**, 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 …

We also present **MacNoise**, a machine-generated, more realistic counterfactual ODQA dataset to provide a more challenging scenario to RALMs

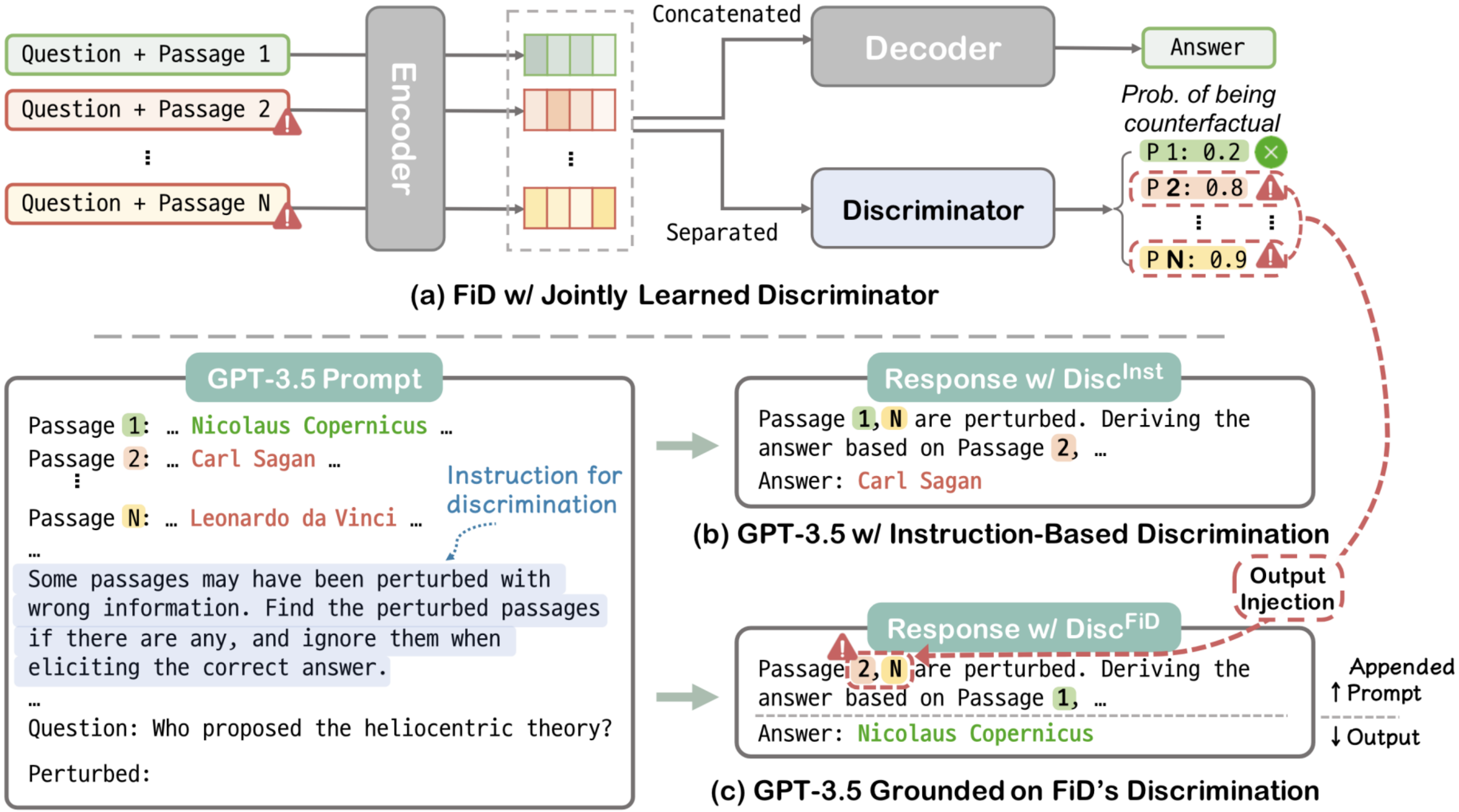
[1] Natural Questions: A Benchmark for Question Answering Research, Kwiatkowski et al., TACL 2019
[2] Entity-based Knowledge Conflicts in Question Answering, Longpre et al., EMNLP 2021

Method: Discern and Answer

We build a framework of RALM called:

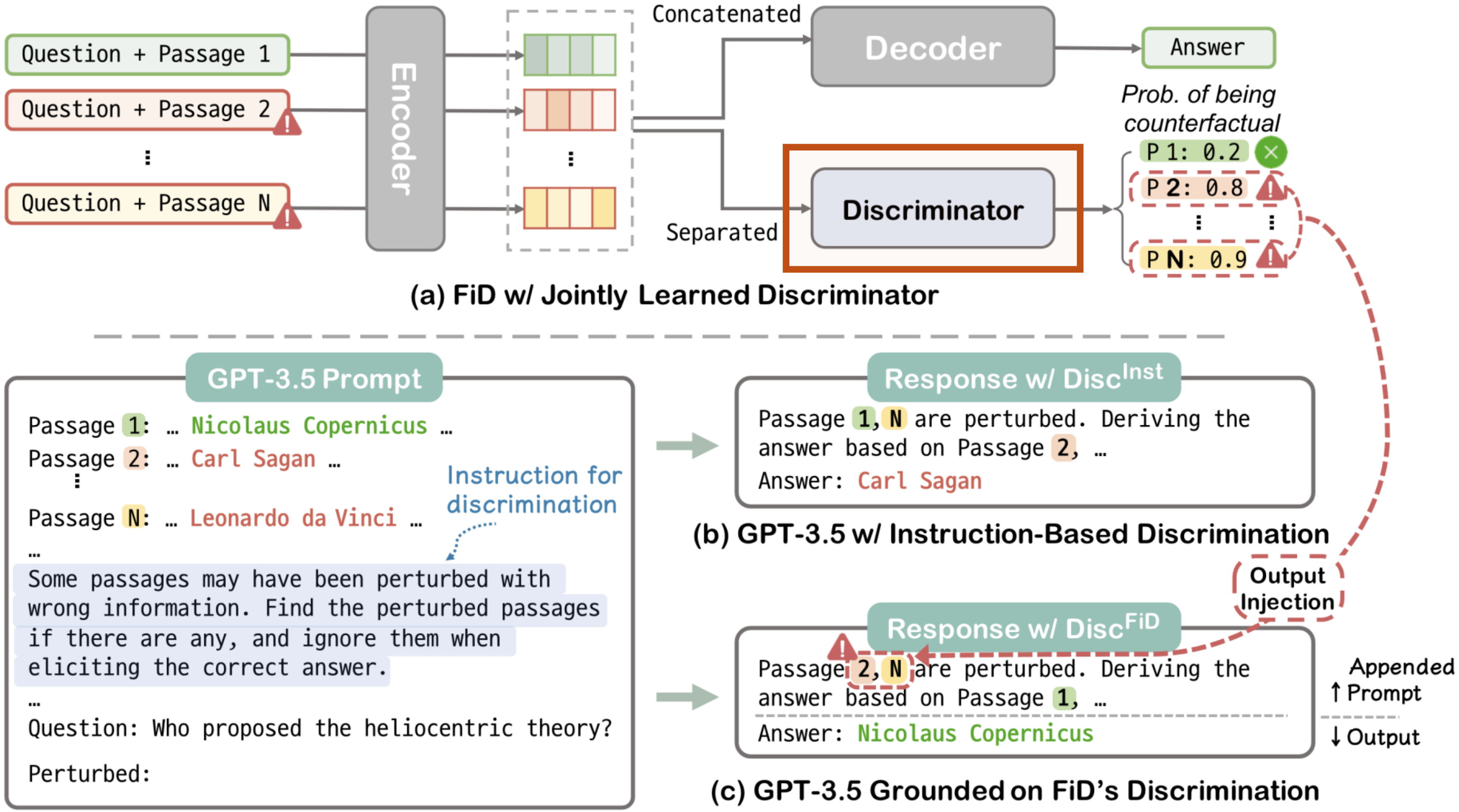
Discern and Answer

Hypothesis: Injecting an inductive bias through the **fine-tuning of a discriminator** enhances LMs ability to "discern" conflicting information



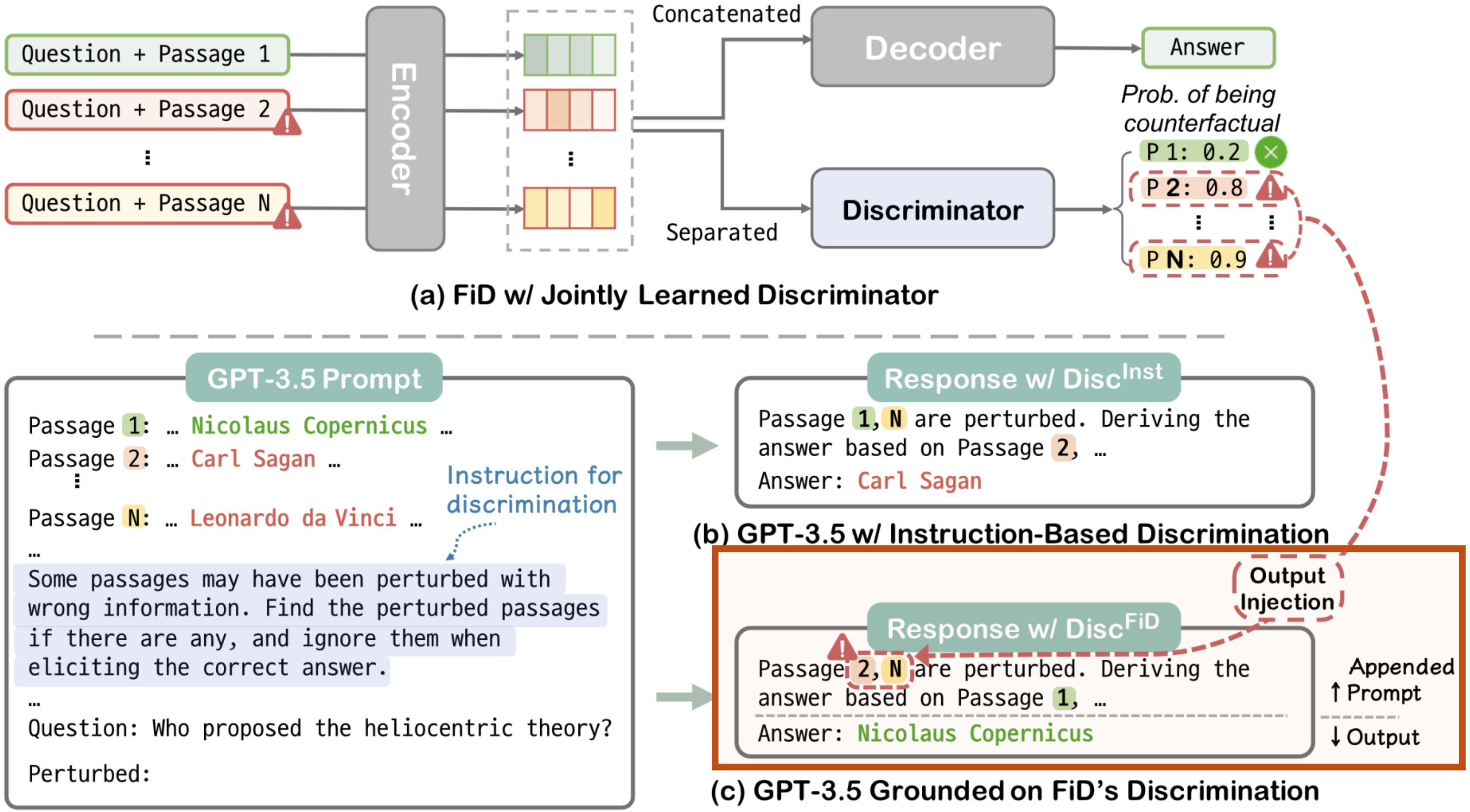
Method: Discern and Answer

Discern and Answer framework builds upon FiD and fine-tunes a **discriminator** that generates a probability of a passage being **counterfactual or not**



Method: Discern and Answer

Discern and Answer framework also **interleaves the high-precision, fine-tuned discriminator outputs with input prompts** for GPT-3.5, leading to improved robustness against noise-injected documents



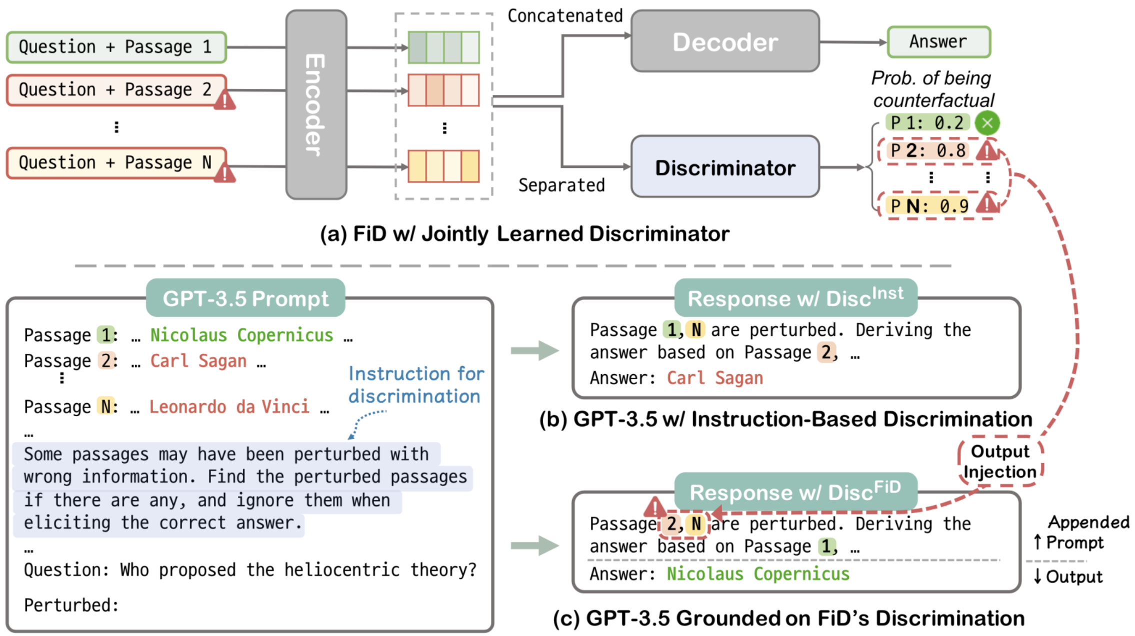
Method: Discern and Answer

Training Objective adopts three loss terms:

$$L_{qa} = -log\, p_{dec}(y|H)$$

$$L_{bce} = \frac{1}{M}\sum_{m=1}^{M} BCE(p_{disc}(t_m|\boldsymbol{h}^{d_m}), t_m)$$

$$L_{contra} = -log \frac{\sum_{d^- \in \mathcal{D}_i^-} exp(p_{disc}(t_m|\boldsymbol{h}^{d^-}))}{\sum_{d^\pm \in \mathcal{D}_i^+ \cup \mathcal{D}_i^-} exp(p_{disc}(t_m|\boldsymbol{h}^{d^\pm}))}$$



L_{qa}: Question-Answering Loss (Auto-regressive loss)

→ Retains the QA ability of the LM

L_{bce}: Binary Cross Entropy Loss

→ Enforces encoder to embed discriminative information in the encoded representations

L_{contra}: Contrastive Loss

→ Jointly considers multiple positives & negatives; prevents overwhelming by the majority class

Experiment Setting: Overview

- **Dataset**
  - Natural Questions (NQ)
  - TriviaQA[2]
- **Document Perturbation Schemes**
  - Entity-Centric (Longpre et al., 2021)
  - Machine-Generated (**MacNoise**)
- **Models**
  - Fusion-in-Decoder (FiD)
  - GPT-3.5 (text-davinci-003)

[1] Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, Joshi et al., ACL 2017

Experiment Setting: Overview

- **Model Settings (FiD and GPT-3.5)**
  - Parametric → Only the base model's parametric knowledge
  - Semi-Parametric → Parametric knowledge + retrieved passages
  - Semi-Parametric + Disc.
    - Disc^{FiD} → Our fine-tuned discriminator for perturbed document detection
    - $\text{Disc}^{\text{Inst}}$ → Discerning through prompt-only method in GPT-3.5

Experiments: Entity Replacement Framework

- **Dataset**
  - Natural Questions (NQ)
  - TriviaQA
- **Document Perturbation Schemes**
  - Entity-Centric (Longpre et al., 2021)

Original Document from Natural Questions (NQ)[1]

… 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** …

⇩

Entity-Centric Perturbation [2]

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

Experiments: Brittleness of RALMs

Base Model	Method	Perturbation % (Dev / Test) 0%	15%	25%	35%	Avg.
FiD	Parametric (w/o Retrieval)	12.1 / 14.7				12.1 / 14.7
	Semi-Parametric	**62.5 / 63.3**	44.5 / 47.7	41.8 / 40.0	28.1 / 30.6	44.2 / 45.4
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**62.5** / 63.2	**51.6 / 51.8**	**43.0 / 45.6**	**38.3 / 36.4**	**48.9 / 49.3**
	Δ Absolute Gain	+0.0 / -0.1	+7.1 / +4.1	+1.2 / +5.6	+10.2 / +5.8	+4.7 / +3.9
GPT-3.5	Parametric (w/o Retrieval)	32.0 / 36.8				32.0 / 36.8
	Semi-Parametric	50.4 / 53.2	40.2 / 45.0	31.3 / 37.8	22.7 / 24.2	36.2 / 40.1
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	48.8 / 54.2	37.9 / 45.6	28.9 / 38.4	21.5 / 26.8	34.3 / 41.3
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**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

Increase in noise among retrieved documents (0% → 35%) leads to

substantially deteriorated performance for both FiD and GPT-3.5

Experiments: Improved Robustness with Discriminators

Base Model	Method	Perturbation % (Dev / Test) 0%	15%	25%	35%	Avg.
FiD	Parametric (w/o Retrieval)	12.1 / 14.7				12.1 / 14.7
	Semi-Parametric	**62.5 / 63.3**	44.5 / 47.7	41.8 / 40.0	28.1 / 30.6	44.2 / 45.4
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**62.5** / 63.2	**51.6 / 51.8**	**43.0 / 45.6**	**38.3 / 36.4**	**48.9 / 49.3**
	Δ Absolute Gain	+0.0 / -0.1	+7.1 / +4.1	+1.2 / +5.6	+10.2 / +5.8	+4.7 / +3.9
GPT-3.5	Parametric (w/o Retrieval)	32.0 / 36.8				32.0 / 36.8
	Semi-Parametric	50.4 / 53.2	40.2 / 45.0	31.3 / 37.8	22.7 / 24.2	36.2 / 40.1
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	48.8 / 54.2	37.9 / 45.6	28.9 / 38.4	21.5 / 26.8	34.3 / 41.3
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**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

Equipping the discriminator **significantly improves robustness** for both FiD and GPT-3.5, especially in settings with high portion of noise (~35%)

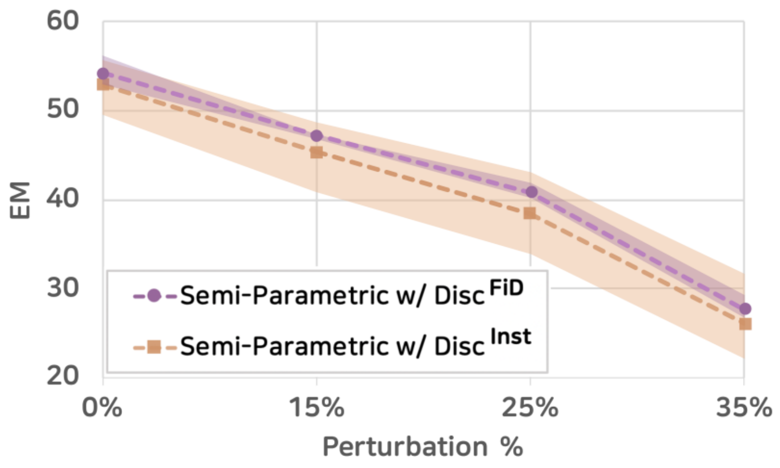
	FiD Prec.	FiD Rec.	FiD F1	GPT-3.5 Prec.	GPT-3.5 Rec.	GPT-3.5 F1
15%	93.49	61.87	74.46	20.98	51.21	29.76
25%	95.77	64.82	77.31	32.32	50.98	39.56
35%	97.14	69.46	81.00	43.42	50.54	46.71

Disc^{FiD} **$\text{Disc}^{\text{Inst}}$**

A **prompt-only discrimination ($\text{Disc}^{\text{Inst}}$)** underperforms **fine-tuned discriminator (Disc^{FiD})** by a large margin

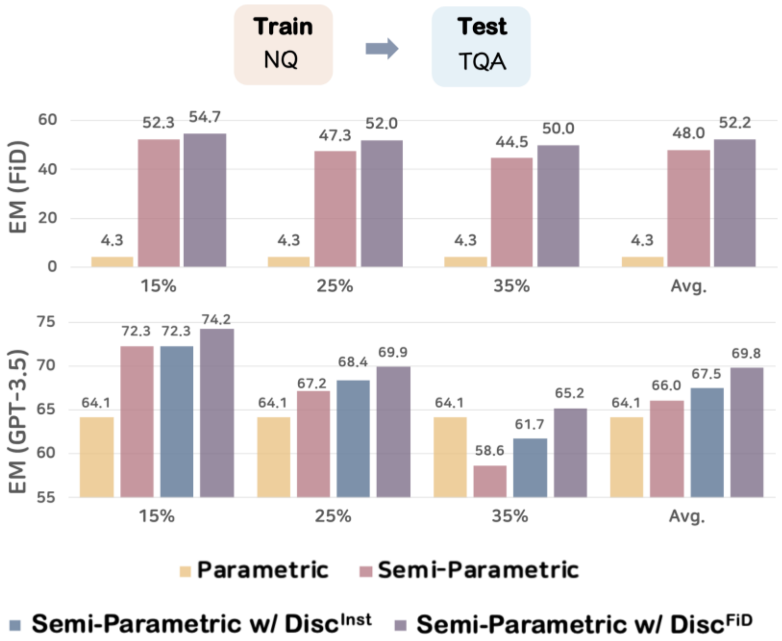
Experiments: Additional Experiments

Enhanced In-Context Learning Stability



In-context learning's stability shows large improvements over GPT-3.5 when interleaved with fine-tuned discriminator output (Disc$^{\mathrm{FiD}}$)

Transferability to TriviaQA



Our results on NQ-open transfers well to TriviaQA, demonstrating the **generalizability of our framework**

MacNoise

- **Limitations of the existing entity-centric perturbation framework (Longpre. et al., 2021)**
  - Context mismatch
  - Confined noise type
  - Semantic equivalence

MacNoise

- **We present MacNoise:**
  - A **Mac**hine-generated **Noise** Dataset for ODQA containing knowledge conflicts among evidence documents
  - Addresses the above limitations of the entity-perturbation scheme
- We use proprietary, SOTA LLMs to generate our documents
  - **GPT-4** : Used to generate our evaluation datasets
  - **GPT-3.5** : Used to generate our training datasets

MacNoise

- **MacNoise constitutes noise-induced passages that retain:**
  - **Question Answerability** – Perturbed passages should **still be answerable given a question**
  - **Length Similarity** – Perturbed passages should be **similar in length to the original document** to avoid any reasoning shortcuts (e.g., length difference)
  - **Answer Perturbation** – Perturbed passages should **not contain the original answer span** or revise the context so that it **no longer supports the answer**

MacNoise

MACNOISE Prompt

```
You are a novel writing AI. Your job is to make up a story based on the following information.
You will be given a question (preceded by "Question:"), a document (preceded by "Document:") and
the corresponding answer ("Answer:"), and you will be asked to create a novel story after ("Revised Document:").
Note, there can be multiple answers (['answer1', 'answer2', ...]) to a given question and document pair.
Now, you should creatively rewrite the document so that the document has a different answer than the given answer(s).

The rewritten document must adhere to all of the following rules:
1) The rewritten document must be answerable by the question.
The information (e.g., entities, phrases) explicitly in the question should not be changed from the original
document.
2) The rewritten document should be similar in length to the given original document above.
3) The rewritten document should not contain the original answer.
If the original answer cannot be removed from the document, rewrite the document so the semantics negate / do not
support the answer.

The following are the possible rewriting strategies:
1) Rewrite the document so the passage no longer supports the answer.
2) Replace the entity in the passage.
3) Negate the sentence the answer span exists so that the original answer span is no longer the answer.
Make sure that the rewritten document is in a completely different style than the original document, and correctly
generate punctuations like periods (".") and commas (",").

You must give your rewritten document only after "Revised Document:".
```

Experiments: MacNoise

- **Dataset**
  - Natural Questions (NQ)
  - TriviaQA
- **Document Perturbation Schemes**
  - **MacNoise** – a new machine-generated knowledge conflict ODQA benchmark

Original Document from Natural Questions (NQ)[1]

… 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 …

Experiments: Brittleness of RALMs

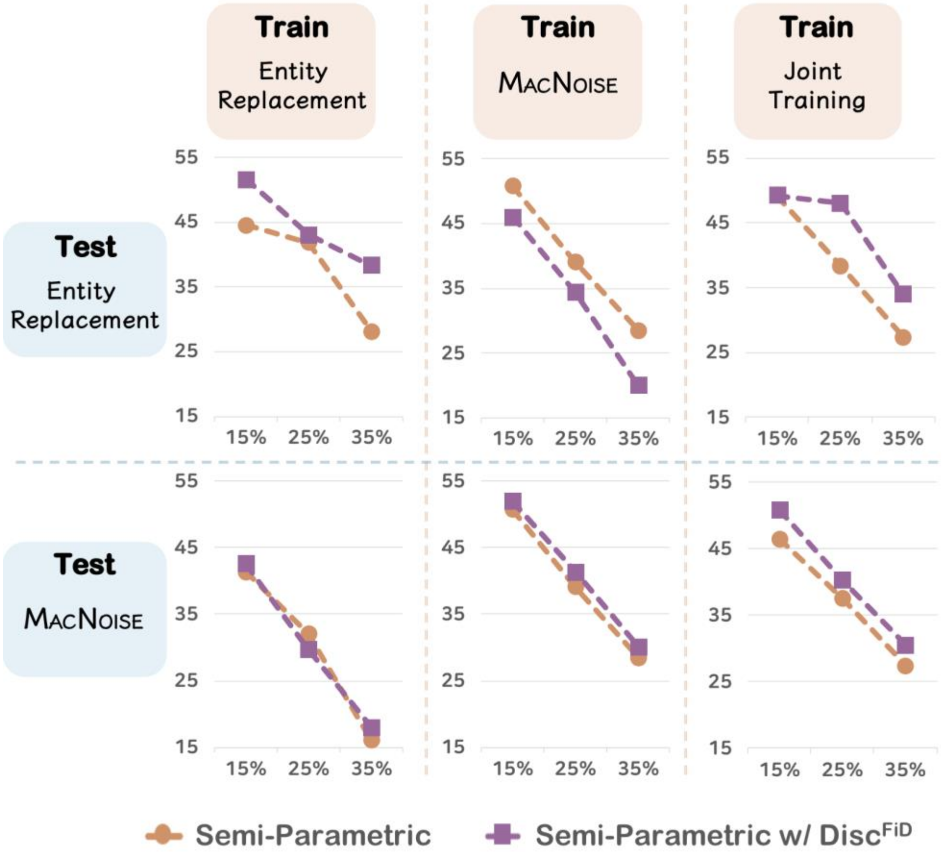
Base Model	Method	Perturbation % (NQ-open) 0%	15%	25%	35%	Avg.	Perturbation % (TQA-open) 0%	15%	25%	35%	Avg.
FiD	Parametric (w/o Retrieval)	12.1				12.1	4.3				4.3
	Semi-Parametric	**62.5**	50.8	39.1	28.5	45.2	**61.7**	54.3	48.8	35.9	50.2
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**62.5**	**52.0**	**41.4**	**30.1**	**46.5**	60.9	**60.6**	**53.5**	**48.1**	**55.8**
	Δ Absolute Gain	+0.0	+1.2	+2.3	+1.6	+1.3	-0.8	+6.3	+4.7	+12.2	+5.6
GPT-3.5	Parametric (w/o Retrieval)	32.0				32.0	64.1				64.1
	Semi-Parametric	50.4	28.5	23.8	16.0	29.7	71.9	60.9	53.5	43.0	57.3
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	48.8	36.3	28.5	19.5	33.3	73.8	64.1	56.6	44.9	59.9
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**51.2**	**37.1**	**30.1**	**21.5**	**35.0**	**76.2**	**68.0**	**61.7**	**53.1**	**64.7**
	Δ Absolute Gain	+0.8	+8.6	+6.3	+5.5	+5.3	+4.3	+7.1	+8.2	+10.1	+7.4

Increase in noise among retrieved documents leads to an even greater drop in MacNoise than in entity-centric perturbation (**34.4 drop from 0% to 35% for MacNoise** vs. 27.7 drop in entity-perturbation)

Experiments: Additional Experiments

After **jointly training our discriminator with the entity-perturbed and MacNoise datasets**, we can see that the discriminator is able to **address the counterfactual noise in both** the entity- and LLM-perturbed settings simultaneously

Complementarity of Entity Perturbation and LLM-generated Noise



Conclusion

- **We propose Discern and Answer**
  - A retrieval-augmented LM framework that addresses the counterfactual information embedded within retrieved documents
- **We build MacNoise**
  - A machine-generated ODQA benchmark that provides a more challenging, realistic setting for RALMs.