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

Giwon Hong[*1] **Jeonghwan Kim**[*2] **Junmo Kang**[*3]
Sung-Hyon Myaeng[4] **Joyce Jiyoung Whang**[†4]
[1]University of Edinburgh [2]UIUC [3]Georgia Tech [4]KAIST
giwon.hong@ed.ac.uk jk100@illinois.edu junmo.kang@gatech.edu
{myaeng, jjwhang}@kaist.ac.kr
*Equal contribution †Corresponding author

Abstract

Most existing retrieval-augmented language models (LMs) assume a naïve dichotomy within a retrieved document set: query-relevance and irrelevance. Our work investigates a more challenging scenario in which even the "relevant" documents may contain misleading or incorrect information, causing conflict among the retrieved documents and thereby negatively influencing model decisions as noise. We observe that existing LMs are highly brittle to the presence of conflicting information in both the fine-tuning and in-context few-shot learning scenarios. We propose approaches for handling knowledge conflicts among retrieved documents by explicitly fine-tuning a discriminator or prompting GPT-3.5 to elicit its discriminative capability. Our empirical results on open-domain QA show that these approaches significantly enhance model robustness. We also provide our findings on incorporating the fine-tuned discriminator's decision into the in-context learning process, proposing a way to exploit the benefits of two disparate learning schemes. Alongside our findings, we provide MACNOISE, a machine-generated, conflict-induced dataset to further encourage research in this direction[1].

1 Introduction

The general framework of retrieval-augmented language models (LMs) for question answering (QA) consists of retrieving documents related to a question using a sparse (Robertson et al., 2009; Jang et al., 2021) or a dense (Karpukhin et al., 2020) retriever, and processing the retrieved documents using encoder (Devlin et al., 2019) or decoder (Raffel et al., 2020) models to derive an answer. Despite being used in many practical applications, most retrieval-augmented LMs (Guu et al., 2020; Lewis et al., 2020; Izacard and Grave, 2021; Lewis et al., 2021) are predicated on a naïve assumption: the retrieved documents are either relevant or irrelevant to the query. However, such a dichotomous view overlooks the fact that in real-world scenarios, the documents purportedly relevant to the query may not consistently offer accurate or reliable information, leading to conflicts among the retrieved documents. Such conflicts, as noise, can adversely affect the models that heavily rely on the veracity of the provided information. Inconsistencies caused by conflicting information may occur for various reasons such as updated/outdated or fabricated/hallucinated information, with the latter being a significantly growing concern due to documents generated by large language models (LLMs) flooding the Web.

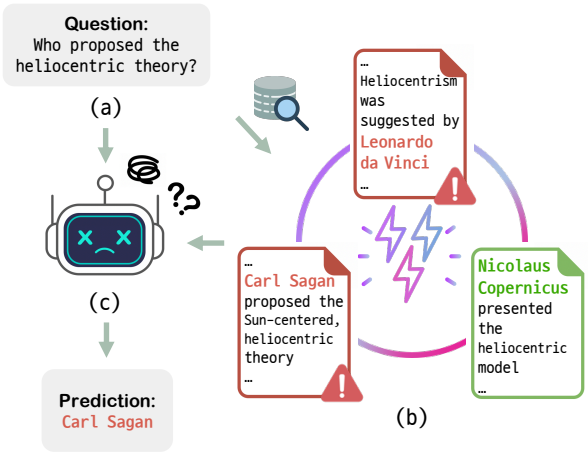

Figure 1: In an ODQA setting, (a) a question is used to retrieve a set of (b) relevant documents which may contain conflict-causing documents that render (c) the retrieval-augmented LMs unreliable.

We study the robustness of retrieval-augmented LMs in the presence of noise and the ensuing knowledge conflict in open-domain question answering (ODQA). To facilitate a controllable study, we adopt the widely-used Longpre et al. (2021)'s framework that deliberately perturbs the retrieved documents, which is also used in previous works

[1]We release our code and dataset at: https://github.com/wjdghks950/Discern-and-Answer

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2474–2495

on knowledge conflict (Chen et al., 2022; Neeman et al., 2023). This deliberate perturbation causes conflict among the documents, as shown in Figure 1, which undermines the model's reliability even in the presence of a gold document.

Our empirical results show existing models such as FiD (Izacard and Grave, 2021) and GPT-3.5 (`text-davinci-003`) (Brown et al., 2020) are highly susceptible to conflicting information. To alleviate this problem, we propose inducing the discrimination capabilities and exploiting them in the fine-tuned (FiD; §3.1) and in-context learned (GPT-3.5; §3.2) models to let them focus on reliable information. We demonstrate that (i) the fine-tuned LM achieves high precision in discerning authentic from counterfactual documents, and (ii) large language models (LLMs) leverage rich parametric knowledge to perform tasks with limited training data, but exhibit weakness in distinguishing noisy documents (§4). Based on our findings, we combine the strengths of fine-tuning and prompting, highlighting the potential benefits of leveraging lightweight fine-tuned LMs to assist LLMs.

Furthermore, while previous works (Chen et al., 2022; Neeman et al., 2023; Si et al., 2023) also leverage Longpre et al. (2021) to emulate knowledge conflict scenarios, the simple entity-swap technique faces several limitations regarding the verisimilitude of the perturbed texts. To this end, we also release a set of LLM-generated contradictory documents using GPT-4 (OpenAI, 2023) to enable a more realistic and challenging study (§5). We hope this can further encourage future works to explore conflict resolution in the retrieval-augmented LMs. Our contributions include:

- We highlight the vulnerability of retrieval-augmented models to counterfactual noise, irrespective of whether they are fine-tuned or in-context learned models.
- We propose a simple yet effective approach for enhancing discrimination capabilities so as to mitigate the model's susceptibility to noise.
- We construct a new LLM-generated counterfactual dataset, MACNOISE, which turns out to be a challenging knowledge-conflict benchmark, as shown in our evaluation.
- Our work opens up a new direction for future works to integrate the benefits of both fine-tuning and in-context learning paradigms.

2 Related Work

Retrieval-Augmented Language Models Retrieval-augmentation aims to capture world knowledge in a more efficient and interpretable manner (Guu et al., 2020), and address the hallucination and knowledge update issues (Lewis et al., 2020; Izacard and Grave, 2021). Some works scaled the size of retrieved documents (Lakhotia et al., 2021), while others adopted retrieval to reduce LM's parameter size (Borgeaud et al., 2022). While promising, most works disregarded the possible prevalence of counterfactual documents. A recent work (Luo et al., 2023) studies instruction-tuned search-augmentation to filter out distracting documents, motivated by the fact that not all retrieved documents are informative. Our work shares a similar motivation but challenges the binary notion of relevance, as even relevant ones can contain incorrect information, causing conflict.

Knowledge Conflicts and Answer Calibration Chen et al. (2022) and Neeman et al. (2023) investigated model behaviors in knowledge conflict settings. They either used calibration (Kamath et al., 2020; Zhang et al., 2021) to abstain from answering, or generated multiple answers upon conflict. Our work, on the contrary, deals with improving the model's ability to distinguish gold from counterfactual information when confronted with knowledge conflicts, providing a correct answer rather than remaining silent. Kazemi et al. (2023) argued that available information is frequently inconsistent or contradictory particularly when reasoning in the real-world. They imposed explicit preferences over information sources to resolve conflicts, whereas our approach aims to modulate models' implicit parametric knowledge through discriminator fine-tuning. A concurrent work, Pan et al. (2023), studies LLM-generated misinformation. While they use GPT-3.5 to generate documents for explicitly distinct settings, we aim for more natural, challenging, and controllable settings using GPT-4, e.g., introducing the controllability of the noise level (§4.1). Our method shows stark contrast to their separate fine-tuning and prompting approaches by explicitly combining the intermediate reasoning steps of prompting with the fine-tuned discriminator to detect misinformation (Figure 2 `(c)`).

Machine-Generated Documents and Misleading Information In recent years, machine-generated documents resembling human-written content have

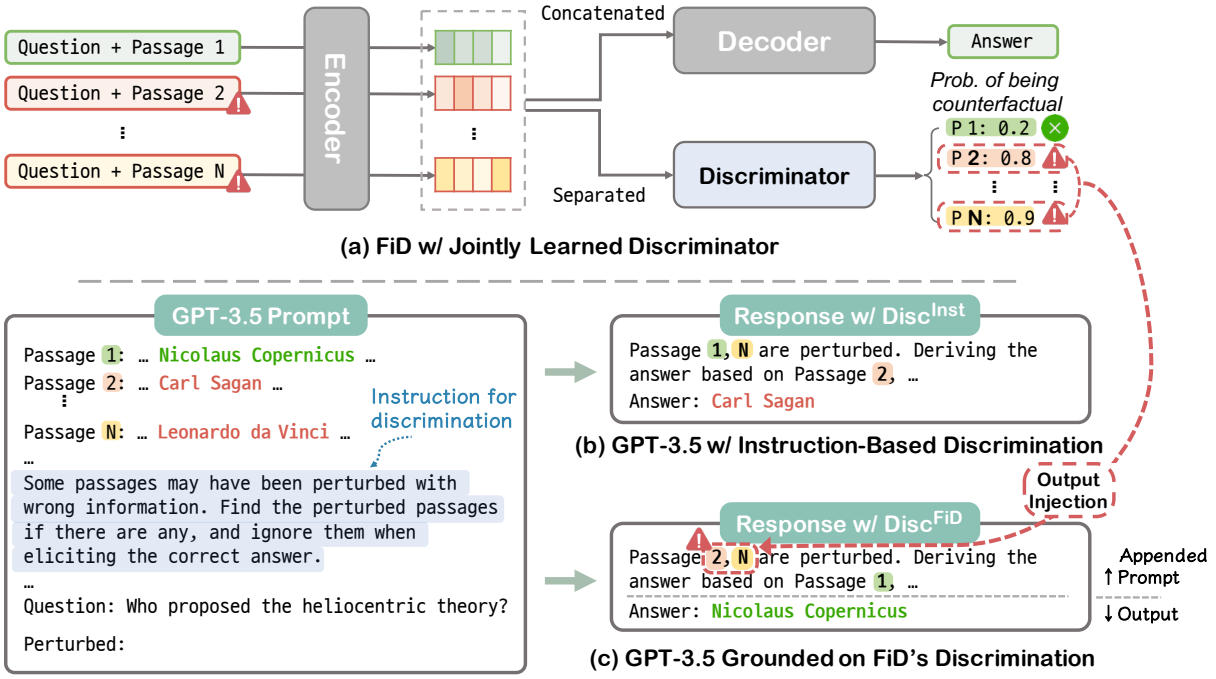

Figure 2: Illustration of our approaches to enhancing robustness to counterfactual noise. (a) Along with the decoder, the discriminator is jointly trained with the downstream task (QA), making the encoder produce corrupt-aware embeddings. (b) GPT-3.5 is prompted to find the perturbed documents before generating an answer. A zero-shot example is shown for brevity. (c) Fine-tuned discriminator output is injected into the prompt for GPT-3.5.

raised concerns about misinformation and differentiating their origins from human-written documents (Ouyang et al., 2022). For instance, recent work has shown that humans struggle to identify machine-generated writing (Clark et al., 2021; Kim et al., 2021b). The emergence of GPT-4 has further intensified worries about the potential misuse of such models to create deceptive content (OpenAI, 2023). Research has revealed that conventional models rarely recognize misinformation but rather contribute to its amplification by generating fabricated details (Zhou et al., 2023). Furthermore, it has been shown that LLM-based applications can be indirectly controlled by adversaries by manipulating retrieval data (Greshake et al., 2023). These studies motivate the need for robust approaches to address the challenges posed by machine-generated documents. Our work contributes to mitigating the influence of such documents, particularly in the context of retrieval-augmented language models for QA.

3 Method: Discern and Answer

We hypothesize that injecting inductive bias (Hong et al., 2022; Kim et al., 2023) about whether a document may be perturbed or not into a retrieval-augmented LM improves model robustness to conflicting information in QA. We equip a QA model with a discriminator learned jointly with a QA task, to interpolate the discriminative features with the encoder embeddings so the decoder can capture such a bias when deriving an answer (Figure 2 (a)). Besides fine-tuning, we explore the potential to elicit GPT-3.5's discriminability through in-context instruction, by letting the model explicitly discern before answering (Figure 2 (b)) or injecting fine-tuned model's output into a prompt (Figure 2 (c)).

3.1 Incorporating Learnable Discriminator into Retrieval-Augmented Model

Our model builds upon FiD (Izacard and Grave, 2021), a retrieval-augmented encoder-decoder LM that leverages DPR (Karpukhin et al., 2020) to retrieve a set of M documents from a text corpus $\{d_1, d_2, ..., d_N\} \in D$, where d_i is retrieved by a similarity search with a question embedding along a document index of size N encoded by a pre-trained BERT (Devlin et al., 2019). Each document d_m is prepended with a question q to be processed independently by a T5 (Raffel et al., 2020) encoder, and is fed to the discriminator (jointly fine-tuned with the encoder). The discriminator is a one-layer feed-forward network that receives as input each

document embedding separately and determines whether the document is perturbed or not; since the information needed to classify a document is encoded by the preceding encoder, a single layer suffices. The intuition underlying our discriminator fine-tuning is to enhance the encoder's ability to navigate its parametric knowledge space. The resulting encoder representations, therefore, are infused with perturbation-indicative latent information that reduces the influence of perturbed documents on the decoder when it attends over them to generate the final answer.

The encoder representations are concatenated (||) along the sequence dimension as follows: $H = ||_{m=1}^{M} Encoder(q, d_m)$, $H \in \mathbb{R}^{M \times T \times E}$, where T is the maximum sequence length per document and E is the embedding size.

The training objective adopts three complementary loss terms: a generative QA loss L_{qa}, a binary cross entropy for discrimination L_{bce}, and a contrastive loss L_{contra}, formulated as follows:

$$L_{qa} = -log\, p_{dec}(y|H) \tag{1}$$

$$L_{bce} = \frac{1}{M} \sum_{m=1}^{M} BCE(p_{disc}(t_m|\boldsymbol{h}^{d_m}), t_m) \tag{2}$$

$$L_{contra} = -log \frac{\sum_{d^- \in \mathcal{D}_i^-} exp(p_{disc}(t_m|\boldsymbol{h}^{d^-}))}{\sum_{d^\pm \in \mathcal{D}_i^+ \cup \mathcal{D}_i^-} exp(p_{disc}(t_m|\boldsymbol{h}^{d^\pm}))} \tag{3}$$

where p_{dec} and p_{disc} denote the decoder and discriminator probability distribution, respectively. y is the ground-truth answer sequence, $\boldsymbol{h}^{d_m} \in H$ is an encoder representation for the m-th document, $t_m \in \{0, 1\}$ is the perturbation label, $\mathcal{D}_i^+$ and $\mathcal{D}_i^-$ are sets of original and perturbed documents, retrieved given the i-th question, respectively. In essence, these three loss components combined ensure a holistic training signal. L_{qa} keeps the primary goal of question answering on track, and L_{bce} retains the encoder's binary classification ability. Inspired by Min et al. (2023), the adopted L_{contra} considers multiple perturbed and original documents, ensuring that the model does not get overwhelmed by the majority class (i.e., original documents) and continues to learn the adequate nuances of perturbed documents via contrastive objective. The final loss is $L = L_{qa} + L_{bce} + L_{contra}$. The effects of each term are discussed in §4.5.

3.2 Instruction-Based Scheme for Enhancing Robustness to Counterfactual Noise

Our work, in addition to fine-tuning, investigates the effectiveness of instructing GPT-3.5 (Ouyang et al., 2022) to figure out the perturbed documents before answering. Our input prompt consists of (i) a set of retrieved documents partly perturbed by our perturbation scheme in §4.1 and §5.2, followed by (ii) a task-specific instruction (Figure 2 (b)) that prompts the model to explicitly identify and ignore the perturbed documents and generate a correct answer, and (iii) the question that follows afterwards (details are in Figure 6 in Appendix C.5).

As an extension, we also incorporate the discriminator (§3.1) to the prompt-based approach. Instead of making GPT-3.5 find the perturbed documents, we insert FiD's discriminator output into the prompt. This way, we combine the GPT-3.5's rich parametric knowledge and the FiD's task-specific discriminator of high precision (Figure 2 (c)), exhibiting complementarity as discussed in §4.3.

4 Evaluation under Entity Replacement Framework (Longpre et al., 2021)

We measure the performance of FiD and GPT-3.5 (`text-davinci-003`) in the following settings. The **Parametric (w/o Retrieval)** setting relies on only rich parametric knowledge (Kim et al., 2022) to answer a question. The **Semi-Parametric** setting uses retrieved documents and parametric knowledge; we measure how the infused conflicting information affects the models' performance. Our methods with discrimination (**`Disc`**) capabilities are denoted as **Semi-Parametric + `Disc`**: the fine-tuned discriminator is superscripted as $\texttt{Disc}^{\texttt{FiD}}$ and the purely prompt-based discrimination as $\texttt{Disc}^{\texttt{Inst}}$.

To fit the maximum length of GPT-3.5, we use the top 5 documents for the dev and test sets for both GPT-3.5 and FiD for a fair comparison. Due to the API budget constraint, we sample 256 dev set as in Le et al. (2022), while using the full test set. The generated outputs from GPT-3.5 are ensembled over the k instances (Appendix D.4) to mitigate the in-context sample sensitivity observed in Zhao et al. (2021). Details are in Appendix C.

4.1 Generating Adversarial Documents

Our study explores *the robustness of models under contradictory information, and the influence of varying degrees of noise*. To facilitate a controllable study, inspired by Kim et al. (2021a), we generate perturbed documents by adopting an entity-centric perturbation strategy (Longpre et al., 2021). This involves taking a document and substituting a

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

Table 1: Performance in Exact Match (EM) on our **dev** and **test** sets (full), according to the perturbation % of retrieved documents. GPT-3.5 is ensembled (Appendix D.4) over $k = 5$ instances (§4). Δ is against Semi-Parametric.

	FiD			GPT-3.5		
	Prec.	Rec.	F1	Prec.	Rec.	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

Table 2: Discriminator performance on our full NQ-Open test set. Each row corresponds to perturbation %.

gold answer with a randomly sampled named entity of the same type, e.g., Michael Jordan (PER) is replaced with Kobe Bryant (PER). We measure the LMs' performance by controlling the proportion of perturbed documents (**0%, 15%, 25%, 35%**). Details about generation are in Appendix B.

4.2 Brittleness of Retrieval-Augmented Models to Conflicting Information

We analyze how brittle the retrieval-augmented LMs are in the presence of conflict-provoking (i.e., perturbed in the experimental setting) documents for the NQ-Open task (Kwiatkowski et al., 2019). In Table 1, we show that the performances of **Semi-Parametric** for both FiD and GPT-3.5 degrade significantly as the perturbation percentage increases, even when the gold documents are provided. We also note that in a highly perturbed setting (**35%**), GPT-3.5's **Semi-Parametric** becomes worse than its **Parametric (w/o Retrieval)** counterpart. Our results demonstrate that these seemingly strong models are easily affected by conflicts.

4.3 Improved Robustness via Discriminators

For FiD, we see that **Semi-Parametric w/** $\texttt{Disc}^{\texttt{FiD}}$ exhibits improved robustness when confronted with conflicting information (**15% - 35%**), with the average gain of 3.9 on test set. As the proportion of misleading noise increases, there is a general drop in performance while our approach, especially in a highly conflicting scenario (e.g., **35%**), exhibits maximum gains. This highlights the discriminator's efficacy in reducing vulnerability to noise.

For GPT-3.5, we observe that $\texttt{Disc}^{\texttt{Inst}}$ does not incite clear improvement. In Table 2, we show $\texttt{Disc}^{\texttt{Inst}}$'s classification performance, where the GPT-3.5's prompt-based few-shot discriminator approach substantially underperforms its fine-tuned counterpart, $\texttt{Disc}^{\texttt{FiD}}$. This motivated us to provide $\texttt{Disc}^{\texttt{FiD}}$'s output to GPT-3.5 as mentioned in §3.2. We find this enhances the LLM's robustness in all degrees of noise, highlighting the synergistic interplay between GPT-3.5's rich parametric knowledge and FiD's precise task-specific discrimination.

We notice that in **35%**, **Semi-Parametric w/** $\texttt{Disc}^{\texttt{FiD}}$ underperforms **Parametric (w/o Retrieval)** despite the performance recovery from **Semi-Parametric** (§4.2). This is attributed to the high portion of noise caused by the suboptimal recall (Table 2), which is exacerbated by GPT-3.5's strong prompt-following characteristics (Ouyang et al., 2022) as evidenced by **Semi-Parametric**'s high susceptibility in §4.2. This indicates room for further improvement in future work.

4.4 Enhanced In-Context Learning Stability

Figure 3 shows the best, average and worst EM scores of GPT-3.5 over 5 different in-context samples. In-context learning is known for its high instability (Zhao et al., 2021; Min et al., 2022), and we discover that injecting the fine-tuned discriminator into the in-context learning (**GPT-3.5 (Semi-parametric w/** $\texttt{Disc}^{\texttt{FiD}}$**)**) greatly improves

Method	QA (EM)				Classification (F1)			
	15%	25%	35%	Avg.	15%	25%	35%	Avg.
Semi-parametric	44.53	41.80	28.12	38.15	-	-	-	-
+ Disc. (L_{qa} + L_{bce})	49.22	43.75	35.94	42.97	73.48	75.77	**82.65**	**77.30**
+ Disc. (L_{qa} + L_{contra})	45.70	**44.92**	37.11	42.58	59.47	71.02	74.91	68.47
+ Disc. (L_{qa} + L_{bce} + L_{contra})	**51.56**	42.97	**38.28**	**44.27**	**74.05**	**77.43**	80.15	77.21

Table 3: Ablation study on the loss terms.

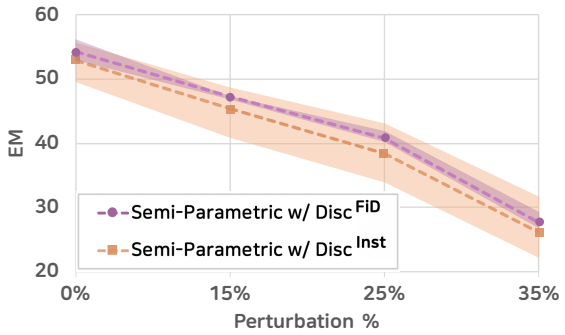

Figure 3: Comparison of GPT-3.5's stability for each discriminator setting. The shaded area represents the variance computed between the best and worst EM.

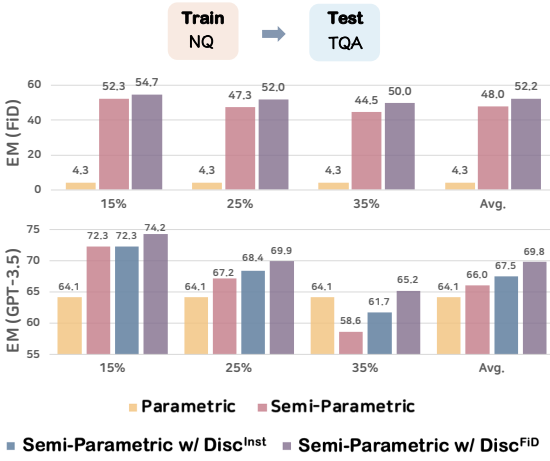

Figure 4: Results on TQA-open **dev**. FiD (i.e., discriminator) is trained on NQ-open and evaluated on TQA-open to examine the transferability of the robustness acquired through our method.

the stability. This new facet along with the result in §4.3, which shows complementarity, highlights the potential of leveraging both strengths of fine-tuning and in-context learning paradigms.

4.5 Ablation Study

To demonstrate the effect of different loss terms in fine-tuning our discriminator, we provide the results of our ablation study in Table 3. The simple binary classification loss, L_{bce}, which is jointly minimized with the QA loss, L_{qa}, markedly improves performance in the perturbed scenarios. We also evaluate the contrastive objective, L_{contra} between perturbed and original documents. While the sole addition of L_{contra} underperforms both the QA and perturbation classification, we show that it shares a complementary relationship with L_{bce}, greatly improving the overall performance across different perturbation configurations; we therefore select this setting as our proposed model.

4.6 Task Transferability to TriviaQA

While our models demonstrate promising results on NQ-open, it remains questionable whether the models can generalize to other datasets. To evaluate the transferability and robustness of our fine-tuned discriminator on other related tasks, we evaluated the performance of our models on the TriviaQA (TQA-open) (Joshi et al., 2017) dev set as shown in Figure 4; the discriminator was fine-tuned only on the NQ-open dataset. The results show that the discriminator is able to distinguish perturbed documents from original ones given the performance gains on the perturbed TQA-open dataset. This suggests that our fine-tuned discriminator, even when it is not explicitly fine-tuned on an end task dataset, is able to extend its discriminability to other tasks. Furthermore, the retention of robustness in the perturbed TQA-open setting serves as a testament that our discriminator does not rely on shortcuts or memorization to distinguish perturbed documents. Test set results exhibiting similar trends are shown in Figure 7 in the Appendix D.1.

5 Evaluation on New Machine-Generated Noise (MACNOISE) Benchmark

To extend our evaluation scope beyond the entity replacement, we present MACNOISE, a **Mac**hine-Generated **Noise** dataset for ODQA containing knowledge conflicts among evidence documents. MACNOISE aims to provide more realistic knowledge conflict scenarios compared to the previous

	Context Mismatch	Question Answerability	Document Length	Counter-factuality	Perturbation Type
Entity Replacement	27.5%	100.0%	106	100.0%	ER (100.0%)
MACNOISE	0.0%	100.0%	123	91.8%	AC (8.9%) | GR (21.9%) | LR (45.2%) | ER (24.0%)

Table 4: Comparison between entity replacement framework vs. our MACNOISE. AC: Additional Context. GR: Global Revision. LR: Local Revision. ER: Entity Replacement w/ Context Match.

entity-centric perturbation framework, addressing limitations discussed in the subsequent section.

5.1 Limitations of Entity Perturbation Framework (Longpre et al., 2021)

While the widely-used entity replacement framework (§4) serves as a simple and scalable proxy for understanding the knowledge conflict scenario in ODQA setting, we posit that, intuitively, the perturbed documents may exhibit the following potential issues:

- Context mismatch: the replaced entities may not be aligned with the co-occurring context (e.g., "Victoria's Secret was founded by *Roy Raymond*, and to his wife Gaye Raymond" to "Victoria's Secret was founded by *Patrick Denham*, and to his wife Gaye Raymond) and this may also entail pronoun mismatch.
- Confined noise type: the perturbation scheme focuses only on removing the existing answer entity from the input passage; it does not employ any other alternative noise generation strategy (e.g., answer negation, multiple answers) that helps enhance the verisimilitude of the documents.
- Semantic equivalence: with low probability, semantically equivalent entities such as aliases may be put in place of the original answer entity within the context (e.g., "The author *Samuel Clemens* wrote 'The Adventures of Tom Sawyer'" to "The author *Mark Twain* wrote 'The Adventures of Tom Sawyer'").

These implausible cases risk the manifestation of shortcuts within the models trained on the entity-swapped documents. As such, we introduce MACNOISE (§5.2) to mitigate the model's reliance on these synthetic cues. Since the three problems with Longpre et al. (2021) may cause robustness issues in our proposed system in a more realistic environment, we fine-tune our **Semi-Parametric + Disc^{FiD}** on LLM-generated knowledge conflict documents (§5.2) from `GPT-3.5-turbo`. To see if our fine-tuned model can fend off a more challenging machine-generated noise among retrieved documents, we generate our evaluation dataset with GPT-4 (OpenAI, 2023), the most powerful existing LLM in both commercial and open-source domains. Our dataset generation's significance is highlighted by the prevalence of machine-generated noise (OpenAI, 2023; Zhou et al., 2023) due to the growing usage of LLMs in general. The notorious hallucination issue inundates the Web environment with noisy, potentially fallacious texts, creating a hazardous environment for retrieval-augmented LMs to exploit knowledge from - another cause for our additional dataset generation. We provide the actual prompt used to generate our dataset using the LLMs in Table 9 in Appendix B.2.

5.2 Generating Counterfactual Documents using Large Language Models

Using GPT-4 and `GPT-3.5-turbo`, we generate our evaluation and training datasets, respectively (dataset generation details in Appendix B.2). To address the limitations of the entity-perturbed documents, we leverage the fact that LLM-generated texts are indistinguishable (Clark et al., 2021) from human-generated texts, and LLMs closely adhere to the given instructions, in which our dataset generation constraints are given. We elaborate on the instruction formulation in this section.

Perturbation Instruction Our perturbation instruction constrains the LLMs with the following rules when generating noise-injected documents: (i) *Question answerability* - perturbed documents should be answerable with the paired question; any information requested by the question can be changed but the documents should retain their relevance to the question. (ii) *Length similarity* - perturbed documents should be similar in length to the original document. We impose this constraint to address GPT-model's notorious tendency to generate verbose texts (Liu et al., 2023). (iii) *Answer Perturbation* - the model should either remove the original answer span or revise the document so the

Base Model	Method	Perturbation % (NQ-open)					Perturbation % (TQA-open)				
		0%	15%	25%	35%	Avg.	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

Table 5: Performance in Exact Match (EM) on our dev of NQ-open and TQA-open w/ machine-generated conflict (MACNOISE), according to the perturbation % of retrieved documents. GPT-3.5 is ensembled (Appendix D.4) over $k = 5$ instances (§4). Δ is against Semi-Parametric.

	FiD			GPT-3.5		
	Prec.	Rec.	F1	Prec.	Rec.	F1
15%	97.58	63.35	76.83	17.72	50.89	25.74
25%	96.57	63.14	76.36	26.13	49.94	34.31
35%	96.32	69.32	80.62	37.94	50.91	43.48

Table 6: Classification performance of our discriminator on the NQ-open with MACNOISE.

context no longer supports the answer.

We also provide the LLMs with a set of revision strategies to create the perturbed documents. The revision strategies are similar to rule (iii), prompting the model to rewrite the document so the document no longer supports the answer, to replace the entities in the passage, or to negate the sentences the answer span appears in so that the original answer span no longer supports the answer. The actual instruction used is described in Appendix B.

Comparison to Entity-Perturbed Documents Here, we provide both the quantitative and qualitative comparison of LLM-generated against entity-perturbed documents; the LLM used here is GPT-4 (OpenAI, 2023). In Table 4, we demonstrate that the LLM-generated documents adequately address the problems of the entity-based perturbation scheme used in §4 while retaining their similarity to the original documents in terms of context length and answer validity rate. We sample 64 instances from the GPT-4-generated dev set of the NQ-open dataset; this consists of a total of 320 documents wherein 146 documents (44.24%) are perturbable.

Through manual analysis on the sampled documents, we identified four perturbation types that distinguish the LLM-generated documents from the entity-perturbed ones: (i) Additional Context - most of the original context is retained while the answers are replaced along with a few additional sentences that justify the replaced entity, which explains the slight increase in **Context Length** in Table 4; (ii) Global Revision - the entire context of a document is largely rewritten by the LLM ; (iii) Local Revision - the original context is largely retained while the answers are replaced with minor edits in the given context; (iv) Entity Replacement w/ Context Match - this is analogous to Longpre et al. (2021) while avoiding context mismatch. Refer to Appendix B for dataset statistics and Table 17 and 18 in Appendix D.5 for case study.

5.3 Brittleness and Enhanced Robustness to LLM-Generated Conflicts

We now benchmark models on our LLM-generated conflicts (MACNOISE). Note that the perturbed documents used for evaluation are generated using a more powerful GPT-4 (OpenAI, 2023), posing a more challenging scenario for our discriminator, which is fine-tuned on a dataset perturbed by `GPT-3.5-turbo`. In Table 5, we note an even greater drop (e.g., 50.4 in 0% → 16.0 in 35% on NQ) for Semi-Parametric GPT-3.5 (`text-davinci-003`) when confronted with our adversarially generated documents, compared to the entity-perturbed ones (50.4 in 0% → 22.7 in 35%) in Table 1. This observation not only exposes the vulnerability of existing models, but also underscores the fact that our MACNOISE benchmark is challenging. Meanwhile, our fine-tuned discriminator enhances the robustness of both models to LLM-generated perturbed documents. In particular, we demonstrate GPT-3.5's over-reliance on the retrieved documents, containing counterfactual texts, can be alleviated to better distinguish perturbed

documents, leading to more accurate answers.

5.4 Complementarity of Entity Replacement and LLM-Generated Perturbations

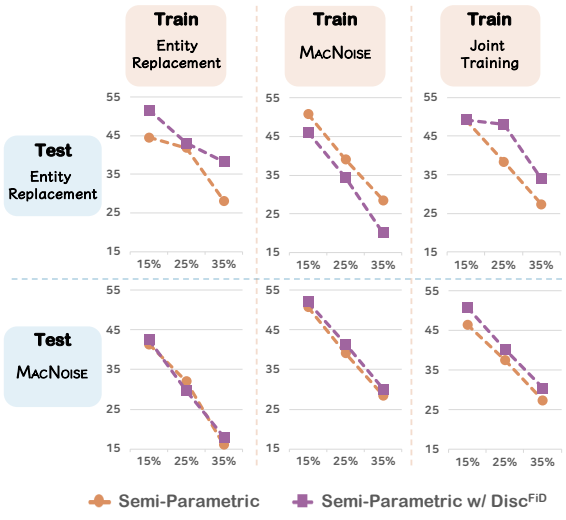

Figure 5: EM scores of the Semi-Parametric and our Semi-Parametric w/ **Disc^{FiD}** on the NQ-open **dev** w/ different perturbations: `Entity Replacement` or MACNOISE. The discriminator is fine-tuned independently (either w/ `Entity Replacement` or MACNOISE) or jointly (`Joint Training`) on the NQ-open **train**.

As an additional experiment, we also evaluate whether the different characteristics of entity-perturbed and LLM-perturbed documents learned during the fine-tuning can be transferred to one another. After jointly training our discriminator with the entity-perturbed (§4) and MACNOISE (§5) datasets, we can see that the discriminator is able to address the counterfactual noise in both the entity- and LLM-perturbed settings simultaneously (Figure 5). This suggests that dealing with different kinds of perturbations simultaneously requires jointly training over the different perturbed document sets, which highlights the importance of curating both the entity- and LLM-perturbed datasets for fine-tuning our discriminator.

6 Conclusion

This work investigates the robustness of retrieval-augmented LMs when the retrieved documents include conflicting information. We show that (i) both the fine-tuned LMs and in-context learned LLMs are brittle to the presence of misleading information, and (ii) our perturbation discriminating approach significantly enhances the LMs' ability to handle conflicts. Furthermore, we find that (iii) combining the fine-tuned discriminator's output with in-context learning improves the LLMs' stability and robustness, creating a new avenue for future work to utilize the advantages of both learning paradigms. We also release MACNOISE, an LLM-generated knowledge conflict dataset for ODQA, to facilitate further research.

Limitations

In the following, we discuss the limitations of our work to encourage future efforts.

Incurred Costs and Data Sampling The use of GPT-3.5 (`text-davinci-003`) for in-context learning (§3.2) incurs substantial cost because of its price ($0.02 per 1,000 tokens). Also, the process of creating our MACNOISE (§5.2) also incurs additional costs because `GPT-3.5-turbo` ($0.002 per 1,000 output tokens) and GPT-4 ($0.06 per 1,000 output tokens) are used to generate our training and evaluation datasets, respectively. To accommodate our budget constraints, we sample 256 instances (Le et al., 2022) from both the NQ-open and TQA-open dev sets. Nonetheless, our results in Tables 1 and 5 clearly demonstrate the efficacy of our proposed fine-tuned discriminator and prompting approaches.

Maximum Input Length of GPT-3.5 Moreover, to fit the maximum input length of GPT-3.5, we use the top 5 documents for the dev and test sets for our baselines GPT-3.5 and FiD to facilitate a fair comparison. We also note that the availability or capability of certain models that need to be accessed through APIs, such as GPT-4 may be subject to change over time.

LLM-Generated Nature of MACNOISE Benchmark MACNOISE is meant to address the synthetic nature of the previous framework (Longpre et al., 2021), in which models may learn to exploit shortcuts to identify misinformation. While emulating more realistic counterfactual documents via MACNOISE, we acknowledge the inherent nature of the LLM-generated data, which can still be deemed synthetic and artifactual (Kang et al., 2020; Hong et al., 2020; Das et al., 2024).

Additional Robustness to Counterfactuals Ideally, our fine-tuned discriminator framework should completely suppress the influence of counterfactual information among retrieved documents for FiD and GPT-3.5. While our method substantially im-

proves the performance of these models with our fine-tuned discriminator when the counterfactual information is present in the retrieved documents, the models are nonetheless influenced by the perturbed documents. We encourage future works to further mitigate the influence of counterfactual information for more robust retrieval-augmented generation in language models.

Ethics Statement

Our work deals with improving the robustness of retrieval-augmented LMs when conflicting information is present among the retrieved documents. To emulate the scenario, our work purposefully, without any ill-intention, perturbed the retrieved documents with the entity-perturbation framework adopted from a previous work (Longpre et al., 2021) and our LLM-generated MACNOISE dataset. Importantly, our goal is to address the issue of misleading information in the ODQA setting. During the validation process for MACNOISE, as presented in Table 4 and elaborated in §5.2, we diligently screened our dataset to ensure the absence of offensive content or personal information. Moreover, given our utilization of GPT-4 for generation, we acknowledge the privacy considerations highlighted in GPT-4 technical report (OpenAI, 2023); The model has been trained on a diverse set of licensed, created, and publicly available data, some of which might encompass publicly available personal information. Nevertheless, stringent steps have been taken to mitigate the potential risks associated with privacy issues.

References

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hung-Ting Chen, Michael JQ Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *arXiv preprint arXiv:2210.13701*.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Debarati Das, Karin De Langis, Anna Martin, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. 2024. Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Giwon Hong, Junmo Kang, Doyeon Lim, and Sung-Hyon Myaeng. 2020. Handling anomalies of synthetic questions in unsupervised question answering. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3441–3448, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Giwon Hong, Jeonghwan Kim, Junmo Kang, and Sung-Hyon Myaeng. 2022. Graph-induced transformers

for efficient multi-hop question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10288–10294, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. 2020. spacy: Industrial-strength natural language processing in python.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Kyoung-Rok Jang, Junmo Kang, Giwon Hong, Sung-Hyon Myaeng, Joohee Park, Taewon Yoon, and Heecheol Seo. 2021. Ultra-high dimensional sparse representations with binarization for efficient text retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1016–1029, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696.

Junmo Kang, Giwon Hong, Haritz Puerto San Roman, and Sung-Hyon Myaeng. 2020. Regularization of distinct strategies for unsupervised question generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3266–3277, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. 2023. Boardgameqa: A dataset for natural language reasoning with contradictory information. *ArXiv*, abs/2306.07934.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2022. Lepus: Prompt-based unsupervised multi-hop reranking for open-domain qa. *arXiv preprint arXiv:2205.12650*.

Jeonghwan Kim, Giwon Hong, Kyung-min Kim, Junmo Kang, and Sung-Hyon Myaeng. 2021a. Have you seen that number? investigating extrapolation in question answering models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7031–7037, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeonghwan Kim, Giwon Hong, Sung-Hyon Myaeng, and Joyce Jiyoung Whang. 2023. FinePrompt: Unveiling the role of finetuned inductive bias on compositional reasoning in GPT-4. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3763–3775, Singapore. Association for Computational Linguistics.

Jeonghwan Kim, Junmo Kang, Kyung-min Kim, Giwon Hong, and Sung-Hyon Myaeng. 2022. Exploiting numerical-contextual knowledge to improve numerical reasoning in question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1811–1821, Seattle, United States. Association for Computational Linguistics.

Jeonghwan Kim, Junmo Kang, Suwon Shin, and Sung-Hyon Myaeng. 2021b. Can you distinguish truthful from fake reviews? user analysis and assistance tool for fake review detection. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 53–59, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Scott Yih, Yashar Mehdad, and Srini Iyer. 2021. FiD-ex: Improving sequence-to-sequence models for extractive rationale generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3712–3727, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nghia Le, Fan Bai, and Alan Ritter. 2022. Few-shot anaphora resolution in scientific protocols via mixtures of in-context experts. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training socially aligned language models in simulated human society.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hongyin Luo, Yung-Sung Chuang, Yuan Gong, Tianhua Zhang, Yoon Kim, Xixin Wu, Danny Fox, Helen Meng, and James Glass. 2023. Sail: Search-augmented instruction learning.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wentau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2023. Nonparametric masked language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2097–2118, Toronto, Canada. Association for Computational Linguistics.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting GPT-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1958–1970, Online. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

A Discussion

A.1 Why Combine GPT-3.5 and FiD?

A crucial inquiry that may arise from our approach is *Why do we need to combine GPT-3.5 and* $\texttt{Disc}^{\texttt{FiD}}$ *despite its worse performance than the FiD counterpart?* Note that our discriminator is easily trainable with our scalable perturbation framework (§4.1). In a low-resource setting, where downstream task instances are scarce, GPT-3.5's few-shot learning capability shines. The lightweight fine-tuned LMs trained on an easily accessible subtask (e.g., perturbation classification) can, therefore, maximize GPT-3.5's capability.

A.2 On Perturbation Schemes

Inconsistencies caused by conflicting information may occur for various reasons such as updated/outdated or fabricated/hallucinated information. Our study explores *the robustness of models under contradictory information, and the influence of varying degrees of noise*. To facilitate a controllable study, we generate perturbed documents by adopting an entity-centric perturbation strategy (Longpre et al., 2021). While Longpre et al. (2021)'s entity perturbation framework has been widely adopted in several previous works (Chen et al., 2022; Neeman et al., 2023), the perturbation framework faces a number of limitations as we elaborate in §5.1. Our work aims to overcome the confines of the entity-only perturbation framework and propose a new perturbation scheme using LLMs, with which we build MACNOISE.

We measure the LMs' performance by explicitly controlling the proportion of perturbed documents (**0%, 15%, 25%, 35%**). The objective of this extensive study using the scalable and controllable framework is that the proportion of misleading information in the real-world is unknown, consistently changes, or varies depending on document sources. We believe that conflicts may potentially occur in other ways as well, but we clarify that exploring those is beyond the scope of our study.

B Generation of Counterfactual Documents to Infuse Conflicting Information

Our work mainly focuses on improving the retrieval-augmented LMs for ODQA when presented with a mixed bag of gold and counterfactual documents.

Split	Total	N/A	PER	ORG	LOC	DATE	NUM
NQ-open							
Train	79,168	27,916	20,136	2,611	4,311	5,343	3,628
Dev	8,757	3,099	2,872	394	461	1,365	566
Test	3,610	1,322	897	139	248	280	165
TQA-open							
Train	78,785	40,252	19,107	5,838	10,141	1,264	2,183
Dev	8,837	4,542	2,120	665	1,123	163	224
Test	11,313	2,891	4,162	2,683	1,017	142	418

Table 7: NQ-Open and TQA-open dataset statistics and the type-wise count on the number of instances perturbed using a substitution framework in Longpre et al. (2021). **N/A** denotes the instances with non-named entity answers that were not perturbed.

Pert. %		# of Pert. Documents	
		NQ-Open	TQA-Open
Dev	**30%**	191 (14.92%)	199 (15.54%)
	50%	312 (24.38%)	317 (24.76%)
	75%	453 (35.39%)	453 (35.39%)
Test	**30%**	1,369 (14.96%)	3,356 (15.03%)
	50%	2,308 (25.22%)	5,572 (24.96%)
	75%	3,471 (37.93%)	7,811 (34.99%)

Table 8: Statistic on the number of documents perturbed from the 256 dev instances and full test instances sampled for our evaluation. Each row represents the perturbation probability, and the **# of Documents** refer to the perturbed documents and their portion in the percentage out of the 1,280 documents for dev (from which 48.05% were perturbable) and the 9,150 documents for test (from which 50.67% were perturbable) in case of NQ-open. For TQA-open, the percentage is calculated based on the 1,280 documents for dev (from which 67.57% were perturbable) and the 14,974 documents for test (from which 67.08% were perturbable).

B.1 Entity Perturbation (Longpre et al., 2021)

The counterfactual documents are generated with the entity-perturbation framework proposed in Longpre et al. (2021)[2]; we use the `corpus-substitution` scheme in this work. While the previous works (Chen et al., 2022; Neeman et al., 2023; Si et al., 2023) use the framework to investigate the effect of knowledge memorization, our work leverages the entity substitution to generate counterfactual documents that contradict what the model has already learned.

We first identify the instances that have the five named entities - PER, ORG, LOC, DATE, and NUM - as their gold answer (as defined in the

[2]https://github.com/apple/ml-knowledge-conflicts/tree/main released under `Copyright (C) 2021 Apple Inc. All Rights Reserved.`

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:".
```

Table 9: A prompt used for counterfactual document generation using large language models.

previous work), and tag each gold named entity answer with a Named-Entity Recognition (NER) tool[3]. Here, we define the "perturbable" documents as those that contain one of the five NER-typed entities as their answers, or non-perturbable otherwise. Then, we use a set of retrieved documents using DPR (Karpukhin et al., 2020), which was provided in the official repository of FiD (Izacard and Grave, 2021)[4] and find the spans in the documents that overlap with the gold answers. We then perturb each document with certain probabilities by substituting every named entity answer with a randomly sampled named entity. To avoid shortcuts and make the perturbed document discrimination task more challenging, we sample from a pool of entities of the same type as the substituted entity, e.g., Michael Jordan (PER) is replaced with Kobe Bryant (PER). Table 7 shows an overview of the NQ-Open dataset (Kwiatkowski et al., 2019) and TQA-Open dataset (Joshi et al., 2017) used in this work and the type-wise number of instances that have named entities as their answers.

To give a detailed overview on the change in the number of documents with the increasing perturbation probability (**Pert. %**), we also provide the perturbed document statistic in Table 8. The statistic elaborates the details about the perturbable documents within the sampled 256 dev set instances and the full test set instances from the NQ-Open dataset in §4. The "full" test set in our work refers to the 1,830 instances from the NQ-Open test set and 4,464 for the TQA-open test set; these are the instances that contain (i) perturbable passages which (ii) lie within the top 5 passages scored by DPR. In generating our training dataset, based on NQ-open, using Longpre et al. (2021), we apply the same aforementioned entity perturbation strategy. For the training details, refer to Appendix C.2

B.2 MACNOISE: Machine-Generated Perturbation

Our MACNOISE dataset also follows the same statistic as the dataset described in the previous Appendix section (§B.1); since only answer-containing documents can be perturbed, meaning both MACNOISE and entity perturbation were applied to the same subset, the statistics are identical to each other. The difference is that the perturbed documents for MACNOISE were generated by

- `GPT-3.5-turbo`: Used to generate the training dataset using NQ-open. This training dataset is part of MACNOISE.
- GPT-4: Used to generate the evaluation dataset for NQ-open and TQA-open. This evaluation dataset is part of MACNOISE.

The instruction prompt template used to generate the perturbed documents in MACNOISE is presented in Table 9. To deal with the extensive cost of generating all the perturbable training documents, we truncate the number of documents perturbed to

[3]We used the spaCy NER tool (version 3.5.1) (Honnibal et al., 2020), an open-sourced natural language processing tool, released under `The MIT License (MIT)`.

[4]https://github.com/facebookresearch/FiD, released under the `Attribution-NonCommercial 4.0 International` license.

20 (T = 20). In the case of building MACNOISE dataset for TQA-open, we added three quality examples to the prompt as in-context demonstrations, where the examples are sampled from the earlier established dataset for NQ-open. This allowed us to ensure consistency in data quality, addressing the issue of OpenAI models that are subject to change over time, which we empirically encountered after the NQ-open creation with MACNOISE.

C Details of Experimental Settings

C.1 Overview

Models In this section, we provide the list of models we used in this work as our baseline for ODQA:

- FiD (Izacard and Grave, 2021): The retrieval-augmented LM used in our experiment. This includes our (i) **FiD (Semi-Parametric w/ $\texttt{Disc}^{\texttt{FiD}}$)** setting in which we fine-tune the discriminator with either the entity-perturbed NQ-open or MACNOISE, and the Semi-Parametric setting.
- GPT-3.5 (Brown et al., 2020): The LLM used in our experiment; the GPT-3.5 model we use as our baseline is `text-davinci-003`. We use the prompts in Figure 6 for evaluation. This includes the **GPT-3.5 (Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$)** setting.

Datasets The datasets we based our perturbation schemes on are as follows:

- Natural Questions (NQ) (Kwiatkowski et al., 2019)[5]: NQ is an English QA dataset consisting of real queries submitted to the Google search engine and Wikipedia documents. We used the open version of the NQ dataset (NQ-open) along with a set of documents retrieved using DPR (Karpukhin et al., 2020). Due to the API budget constraint, we sample 256 dev set as in Le et al. (2022), while using a full test set of 1,830 instances. There are a total of 79,168 training instances, 8,757 dev instances (from which 256 are sampled due to API budget constraint), and 3,610 test instances (from which 1,830 instances were perturbable). We provide additional details about generating the training and evaluation dataset in Appendices B.1 and B.2.
- TriviaQA (TQA) (Joshi et al., 2017)[6]: TQA is another English-oriented QA dataset, featuring queries sourced from a collection of 14 trivia and quiz-league websites. Specifically, we used the open, unfiltered version of TQA-open, akin to the process with NQ-open, retrieving documents from Wikipedia using DPR as in FiD (Lakhotia et al., 2021). Due to the API budget constraint, we sample 256 instances from the dev set, which consists of a total of 8,837.

Perturbation Schemes In this section, we provide a list of the entity perturbation schemes we used to perturb the datasets:

- Entity Perturbation (Longpre et al., 2021): This method involves the direct replacement of one target entity with another random entity of the same type. The details of generating this dataset are provided in Appendix B.1.
- MACNOISE: Our new machine-generated noise dataset created by `GPT-3.5-turbo` (for training dataset) and GPT-4 (for evaluation dataset) using the prompt given in Table 9. For additional details, refer to Appendix B.2.

C.2 Settings of FiD-based Models

FiD[7] used in this work was based on T5-base (220M parameters) and trained to use a fewer number of retrieved passages due to our computing resource constraints. While FiD's base setting uses T = 100 retrieved passages to answer open-domain questions, our work only considers T = 50 for the entity perturbation scheme (Longpre et al., 2021) and T = 20 for MACNOISE. This, however, does not present an issue to our study, since the findings in Chen et al. (2022) show that FiD tends to focus its attention on the top N, where $N \leq 20$, retrieved documents when generating an answer. For model training, the perturbable probabilities were set to 30%, 50%, and 75% to match the dev/test sets perturbed portions 15%, 25%, 35%, respectively (Table 8). During training, every document undergoes random perturbation based on set probabilities, unlike during evaluation where perturbations are predefined. Our model was fine-tuned in the above settings independently. This fine-tuned model was

[5]The dataset is released under the `Creative Commons Share-Alike 3.0` license.

[6]The dataset is released under the `Apache 2.0` license.

[7]The models were trained using GeForce RTX3090 (24GB VRAM), AMD Ryzen Threadripper 3960X, and 128GiB RAM. The model training took approximately 80 hours.

Hyperparameter	FiD
Batch size	1
Gradient Accumulation	64
Hidden size	768
Max. Sequence length	200
Learning rate	1e-4
Optimizer	AdamW
Seed	42

Table 10: Hyperparameters of the fine-tuned FiD in this work. We set the gradient accumulation to 64 to account for the batch size in the original FiD (Izacard and Grave, 2021). Each passage is of Max. Sequence Length.

Hyperparameter	GPT-3.5	ChatGPT	GPT-4
Context length	4,097	4,097	8,192
top_p	1.0	1.0	1.0
temperature	0.0	0.0	0.0
logprobs	10	N/A	N/A

Table 11: Hyperparemters of the GPT-3.5 (`text-davinci-003`), ChatGPT (`GPT-3.5-turbo`), and GPT-4 used in our experiments.

used in our experiments throughout. We believe that this setting is valid because in real life, we can sample from the real-world Web and identify the sampled distribution of misleading information.

We also provide the important hyperparameters used to train our **FiD (Semi-Parametric w/ Disc$^{\text{FiD}}$)** model (Table 10). For all the other settings, including the size of the training dataset and the gradient steps, we follow the settings specified in the original FiD. Since our experiments demonstrated clear results sufficiently to validate our hypothesis made in this work, we did not perform a hyperparameter search, and the models were trained once with a fixed seed.

C.3 Settings of Large Language Models

Large Language Models (LLMs) used in this work are twofold: our baseline for ODQA[8] (text-davinci-003) and perturbation sources for our MACNOISE dataset (`GPT-3.5-turbo` and GPT-4). We use the aforementioned LLMs through black-box API calls, and we provide the hyperparameters we used in API requests in Table 11. We set `logprobs` as 10 for GPT-3.5 (`GPT-3.5-turbo`) to get top-10 generated answers for the ensemble strategy described in Appendix D.4. For prompt designs, refer to Appendices B.2 (ODQA baseline) and C.5 (generating MACNOISE dataset). We set the number of documents used during evaluation to $T = 5$, since the context window of GPT-3.5 is limited.

[8] A total cost of approximately $5,500 was incurred for API usage for ODQA experiments.

C.4 Joint Training on MACNOISE and Longpre et al. (2021)

As discussed in §5.4, our work further investigates the transferability and complementarity of the entity-perturbed and LLM-perturbed datasets in an effort to address both perturbation schemes with our fine-tuned discriminator model. We jointly fine-tune the **FiD (Semi-Parametric w/ Disc$^{\text{FiD}}$)** model with both the entity-perturbed and LLM-perturbed NQ-open datasets by simply aggregating the two datasets together to form a joint training dataset as a whole. Here, we use the same number of documents ($T = 20$) as the models for MACNOISE do to make the resulting data balanced in terms of the perturbation type.

C.5 LLM Prompt Designs for ODQA

In Figure 6, we explain in detail the design of our prompts used for ODQA. We divide the prompts into four discrete categories, with each one representing one of the four model settings in §4. Following the findings in Khalifa et al. (2022), we place the instruction prompt ("*Refer to the above documents and your knowledge ...*") after the retrieved documents, which takes the advantage of the *recency bias* phenomenon evidenced in a previous work (Zhao et al., 2021). The retrieved documents that precede the instruction are from the k in-context instances ($k = 5$) sampled from the held-out set; the held-out set refers to the remaining dev set instances aside from the 256 randomly sampled dev set used in our experiments. To maximize the effect of our ensembling strategy, we sample the k instances so each has a unique answer NER type and a different number of perturbed documents. We then ensemble (Min et al., 2022; Le et al., 2022) over k separate one-shot iterations for a single test instance to mitigate the in-context sample sensitivity observed in Zhao et al. (2021). Our approach ensembles over the k iterations by marginalizing over the probability of the top 10 generated answers and chooses an answer with the maximum probability (Refer to Appendix D.4). Following the prompt is the one-shot in-context QA pair that guides the model to generate an appropriate answer given a set of retrieved documents and a question. The **Perturbed:** prompt that follows the question and GPT-3.5's generated response enables the model to

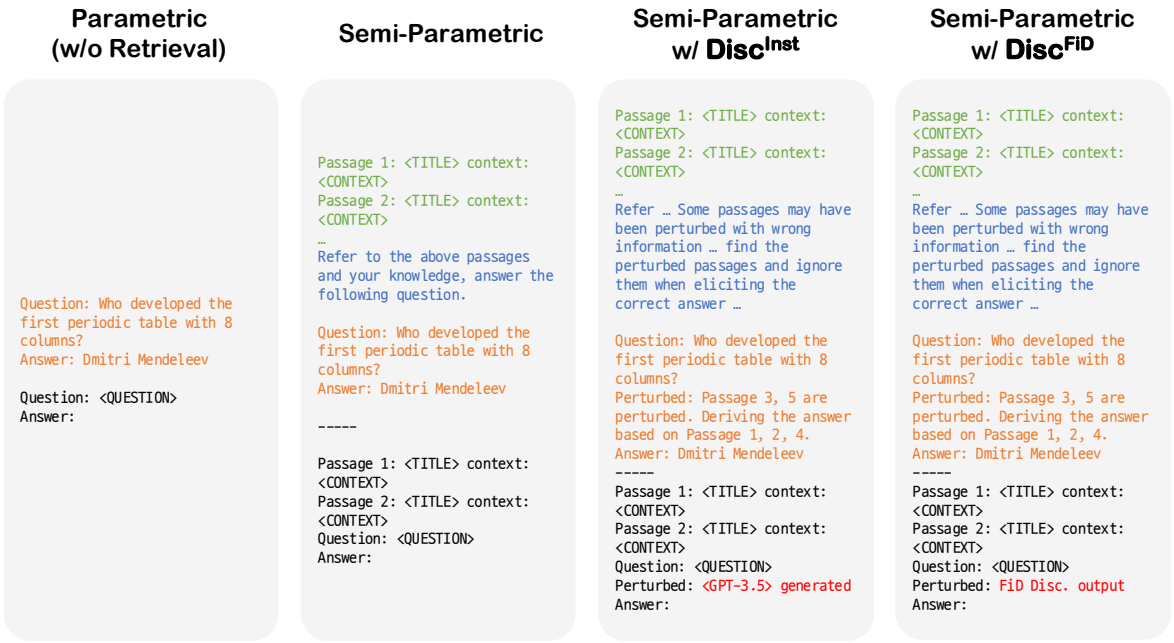

Figure 6: The prompt variants used in our experiments to evaluate the robustness of GPT-3.5 `text-davinci-003` when given a mixed bag of conflicting information-infused documents. The text in **orange** refers to the in-context QA sample from the training data, **green** refers to the in-context retrieved document that corresponds to the **QA pair**, **blue** refers to the prompt, and **red** refers to either the GPT-3.5 generated perturbation classification, or the output of the FiD's discriminator fed straight into the prompt. The **in-context sample documents** may or may not be perturbed. The text in black refers to the evaluation instance.

	FiD			GPT-3.5		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
15%	93.60	61.26	74.05	20.14	49.11	22.67
25%	98.51	63.78	77.43	30.29	48.59	37.32
35%	96.28	68.65	80.15	42.03	49.14	45.31

Table 12: Classification performance of our discriminator on the sampled entity-perturbed NQ-open set (256 instances). Each row corresponds to perturbation %.

	FiD			GPT-3.5		
	Prec.	**Rec.**	**F1**	**Prec.**	**Rec.**	**F1**
15%	80.42	57.79	67.25	16.94	48.34	25.09
25%	90.20	58.04	70.63	24.85	48.08	32.76
35%	94.35	58.94	72.55	39.89	51.21	44.85

Table 13: Classification performance of our discriminator on the sampled entity-perturbed TQA-open dev set.

discern perturbed from original documents. Note that what comes after the **Perturbed:** can be explicitly replaced with the FiD's jointly trained **Disc^{FiD}** output.

D Additional Experimental Results

D.1 Additional Results on Entity Perturbation

Classification In Table 12 and Table 13, we provide the performance of our discriminator on our sampled NQ-Open and TQA-open dev sets, respectively. The FiD result shows that the discriminator classifies perturbed and original documents with high precision, while recall lags behind.

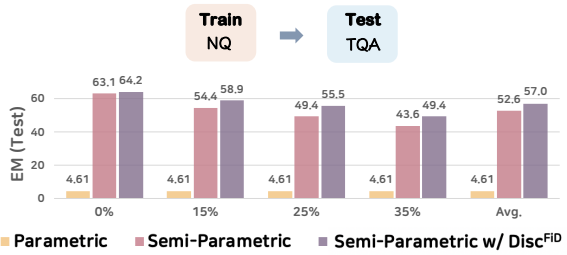

Figure 7: Results of FiD-based models on TQA-open (**test**). Models are trained on NQ-open and evaluated on TQA-open to examine the transferability of the robustness acquired through our method.

Transferability to TQA-open We also demonstrate in Figure 7 the transferability of our NQ-open fine-tuned **FiD (Semi-Parametric w/ Disc^{FiD})** to TQA-open **test** dataset. The results demonstrate

	FiD			GPT-3.5		
	Prec.	Rec.	F1	Prec.	Rec.	F1
15%	94.32	41.71	57.84	14.89	51.46	23.09
25%	93.98	49.21	64.60	23.36	52.24	32.28
35%	93.77	53.20	67.89	36.46	54.30	43.63

Table 14: Classification performance of our discriminator on the sampled TQA-open dev set with MACNOISE.

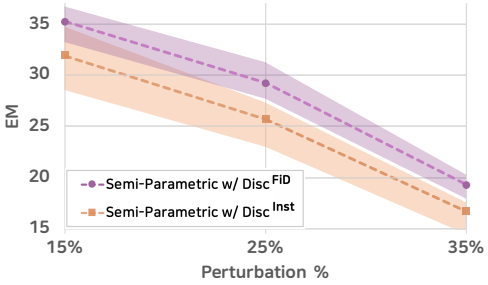

Figure 8: Comparison of GPT-3.5's stability for each discriminator setting on MACNOISE. The shaded area represents the variance computed between the best and worst EM.

similar trends as those shown in §4.6.

D.2 Additional Results on MACNOISE

Classification In Table 14, we provide the performance of our discriminator on our MACNOISE TQA-open dev set. The FiD result shows that the discriminator classifies perturbed and original documents with high precision, while recall lags behind.

Enhanced Stability in GPT-3.5 Injecting the output decisions of our fine-tuned discriminator on MACNOISE into GPT-3.5's prompts, as shown in Figure 8, notably improves the stability of the LLM prediction. Similar to the result in §4.4, the dotted lines in Figure 8 represent the average values over the ensemble, and the top and bottom of the shaded regions represent the worst and best cases, respectively.

D.3 Qualitative Analysis on the Cross-Attention Weights of FiD models

To investigate the effect of the learned discriminator on the answer generation by distinguishing perturbed from original entities, we conduct a qualitative study on the cross attentions of the samples shown in Figure 9. The blue lines visualized[9] denote the attention weight from the last layer of the decoder (i.e., starting token) to the encoder's output representations (i.e., input documents). In the first case, (a) shows that given a counterfactual entity, `Perez Hilton`, the **FiD (Semi-Parametric)** setting does not prevent the decoder from attending to the perturbed entity, neglecting the original entity, `Gorsuch`. On the contrary, in (b), our **FiD (Semi-Parametric w/ $\texttt{Disc}^{\texttt{FiD}}$)** setting, the decoder successfully attends to the original entity, `Gorsuch`, even in the presence of the perturbed entity. We also provide an additional before and after case in (c) and (d), where the original entity, `Steven Weber` is replaced by `Blair Walsh`. In (c), we show that **FiD (Semi-Parametric)** strongly attends to `Blair Walsh`, the perturbed entity, even in the presence of the two original entity spans in the given context. With our discriminator, we show in (d) that the model now attends to the two original entity spans correctly, successfully neglecting the perturbed entity. These cases serve as a testament that our learned discriminator enables the model to effectively control its attention from context-irrelevant, counterfactual entity to the original entity.

[9]The weights were visualized using BertViz (Vig, 2019).

	Ensemble	Average
0%	**51.17**	49.14
15%	**42.18**	40.70
25%	**33.98**	33.90
35%	**27.34**	25.78

Table 15: Comparison between the ensemble of the top-10 probabilities of the generated answer over k = 5 iterations and an average of output scores over the iterations for Semi-Parametric w/ $\texttt{Disc}^{\texttt{FiD}}$.

D.4 Ensemble Strategy in GPT-3.5

In the Experiments (§4 and Appendix C.5), we explain our use of ensemble strategy over the k iterations and the marginalization over the top-10 generated answers to choose our final answer. One notable phenomenon spotted during our experiments is the ensemble's effect of improving over the simple average baseline (Table 15).

The ensemble strategy consistently outperforms the average setting across varying degrees of conflicting information. This suggests that not only does the ensemble of GPT-3.5 outputs alleviate the notorious sample variance issue, but it also enables the model to consider more probable output tokens across the iterations by avoiding the maximum likelihood outputs. One thing we would like to note is that our ensemble strategy demonstrates consis-

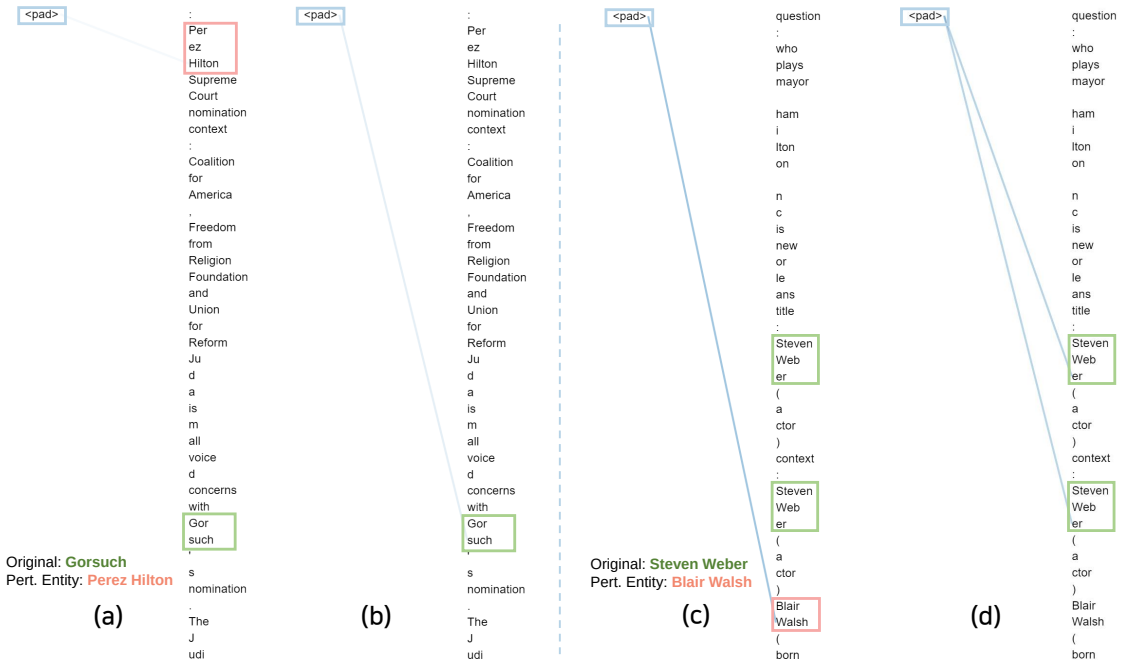

Figure 9: Illustration of our qualitative case study on the cross-attention weights. (a), (b) and (c), (d) are the before (**FiD (Semi-Parametric)**) and after (**FiD (Semi-Parametric w/ $\texttt{Disc}^{\texttt{FiD}}$)**) screenshots of our attention scores, respectively. The perturbed entities are represented in **red** and the original entities are represented in **green**. The <pad> tokens are the starting input token to the FiD decoder and the lines denote the decoder's last layer cross attention to the encoder's output representations (represented here as input sequences).

tent patterns across various configurations (i.e., the number of samples) as shown in Table 16.

D.5 Case Study on Perturbations

In Table 17 and Table 18, we present side-by-side examples of documents of MACNOISE against those of entity perturbation (Longpre et al., 2021). The comparison spans various perturbation types, namely Global Revision, Local Revision, Additional Context, and Entity Replacement w/ Context Match.

Global Revision. We can see that MACNOISE significantly restructures and updates the document's context to provide a more contemporary account. Specifically, it updates the narrative to reflect the events and performance of the Buffalo Bills during the 2020 season. This approach is comprehensive, ensuring the primary theme—how the Buffalo Bills performed during a particular season—remains consistent, but the details and timeline are considerably different. On the other hand, the Entity Replacement method, opts for a very specific and dramatic alteration. By replacing the year "1995" with "between 1652 and 1674," the document becomes factually incorrect but unnatural.

Local Revision. Observations indicate that MACNOISE entails nuanced changes tailored to fit an introduced narrative, while preserving the overarching theme. The founder of Victoria's Secret, originally "Roy Raymond," morphs into "John Thompson," with the surrounding context adjusted for coherence. While fundamental elements like the inception date and brand inspiration remain unchanged, specifics like names get modified. The Entity Replacement technique, in contrast, directly swaps "Roy Raymond" with "Patrick Denham," retaining the majority of the original narrative, which can result in potential mismatches. For example, the unchanged last name of Roy Raymond's wife might cause confusion.

Additional Context. It becomes evident that the perturbation introduced by MACNOISE provides an extended narrative, integrating not only changes in key entity details but also furnishing supplementary information that was not present in the original document. This seems to enrich the content, thereby providing more context, which makes it more realistic and challenging. For instance, while the original narrative emphasizes Joe Spano's acting journey, the MACNOISE perturbed version broadens the discourse, introducing Michael Thomas Grant's

Number of Samples	Method	Perturbation % (Dev)				
		0%	15%	25%	35%	Avg.
$k = 1$	Parametric (w/o Retrieval)		33.66			33.66
	Semi-Parametric	50.22	42.95	35.25	22.24	37.67
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	52.73	45.96	38.91	26.12	40.93
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**53.88**	**46.94**	**40.82**	**27.54**	**42.30**
$k = 2$	Parametric (w/o Retrieval)		35.90			35.90
	Semi-Parametric	51.97	44.10	36.99	23.33	39.10
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	54.48	47.05	39.78	27.10	42.10
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**55.36**	**48.85**	**42.35**	**28.52**	**43.77**
$k = 3$	Parametric (w/o Retrieval)		36.50			36.50
	Semi-Parametric	52.40	45.25	38.09	24.43	40.04
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	55.36	46.78	39.67	26.67	42.12
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**56.50**	**49.23**	**42.46**	**28.42**	**44.15**
$k = 4$	Parametric (w/o Retrieval)		36.50			36.50
	Semi-Parametric	52.68	45.03	36.78	23.11	39.40
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	55.03	46.34	39.95	27.54	42.21
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**56.45**	**49.18**	**41.86**	**29.07**	**44.14**
$k = 5$	Parametric (w/o Retrieval)		36.83			36.83
	Semi-Parametric	53.17	44.97	37.76	24.21	40.03
	Semi-Parametric w/ $\texttt{Disc}^{\texttt{Inst}}$	54.19	45.63	38.41	26.78	41.26
	Semi-parametric w/ $\texttt{Disc}^{\texttt{FiD}}$	**56.28**	**49.18**	**41.64**	**28.63**	**43.93**

Table 16: GPT-3.5 results for ensembling over a different number of samples (k is the number of in-context samples). Performance is reported in Exact Match (EM) on our entity-perturbed NQ-open **dev** set, according to the perturbation % of retrieved documents.

multifaceted talents and achievements. Conversely, the Entity Replacement strategy simply swaps "Joe Spano" for "Jaeden Lieberher," leaving the bulk of the content unaltered, which can lead to contextual mismatches.

Entity Replacement w/ Context Match. Longpre et al. (2021), directly replaces the name of the original entity, "Middle Island," with an unrelated entity, "Mid Glamorgan." This alteration results in a direct substitution without modifying the surrounding context, which can create inconsistencies. For instance, "Mid Glamorgan" does not correspond to any known location within Lake Erie, Ontario. Such discrepancies can make the document unrealistic, and not challenging to discern. On the other hand, MACNOISE also attempts an Entity Replacement but offers a slightly different name, "South Point Island," instead of "Middle Island." This change, while still altering the main entity, remains consistent with the rest of the context, providing a plausible and confounding alternative to the original entity.

Question
when was the last time the bills won their division

Answer
1995

Original Document
title: 1995 Buffalo Bills season context: 1995 Buffalo Bills season The 1995 Buffalo Bills season was the 36th season for the club and its 26th in the National Football League. After suffering a losing season the previous season, the 1995 Bills won the AFC East, and made the playoffs for the seventh time in eight seasons. They defeated the Miami Dolphins in the Wild Card Game in Don Shula's final game as an NFL head coach. The following week they lost to the eventual AFC Champions, the Pittsburgh Steelers. The playoff win over Miami represents the last time to date that the Bills have won a

Entity Replacement (Longpre et al., 2021)
title: between 1652 and 1674 Buffalo Bills season context: between 1652 and 1674 Buffalo Bills season The between 1652 and 1674 Buffalo Bills season was the 36th season for the club and its 26th in the National Football League. After suffering a losing season the previous season, the between 1652 and 1674 Bills won the AFC East, and made the playoffs for the seventh time in eight seasons. They defeated the Miami Dolphins in the Wild Card Game in Don Shula's final game as an NFL head coach. The following week they lost to the eventual AFC Champions, the Pittsburgh Steelers. The playoff win over Miami represents the last time to date that the Bills have won a

MacNoise (Perturbation Type: Global Revision)
title: 1995 Buffalo Bills season context: The 2020 Buffalo Bills season marked the 61st season for the club and its 51st in the National Football League Under the guidance of head coach Sean McDermott, the Bills improved on their 10-6 record from 2019 and won their first AFC East division title since 1995, finishing with a 13-3 record. They made the playoffs for the seventh time in eight seasons and won their first playoff game since 1995, defeating the Indianapolis Colts in the Wild Card round. The Bills continued their playoff run by defeating the Baltimore Ravens in the Divisional round but ultimately lost to the Kansas City. Chiefs in the AFC Championship Game The 2020 season represents the last time to date that the Bills have won their division.

Question
who was the creator of victoria's secret

Answer
Roy Raymond

Original Document
title: Victoria's Secret context: Victoria's Secret Victoria's Secret is an American designer, manufacturer, and marketer of women's lingerie, womenswear, and beauty products. Founded in 1977 as a response to packaged underwear, which the company's founder considered to be "ugly, floral-print nylon nightgowns", 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, in San Francisco, California, on June 12, 1977. Eight years prior to founding Victoria's Secret, in the late 1960s, Raymond was embarrassed when purchasing lingerie for his wife at a department store. Newsweek reported him looking back on the

Entity Replacement (Longpre et al., 2021)
title: Victoria's Secret context: Victoria's Secret Victoria's Secret is an American designer, manufacturer, and marketer of women's lingerie, womenswear, and beauty products. Founded in 1977 as a response to packaged underwear, which the company's founder considered to be "ugly, floral-print nylon nightgowns", 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, in San Francisco, California, on June 12, 1977. Eight years prior to founding Victoria's Secret, in the late 1960s, Raymond was embarrassed when purchasing lingerie for his wife at a department store. Newsweek reported him looking back on the

MacNoise (Perturbation Type: Local Revision)
title: Victoria's Secret context: Victoria's Secret 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. The idea for Victoria's Secret was born when John felt embarrassed while purchasing lingerie for his wife at a department store in the late 1960s. Recognizing the need for a more welcoming and specialized store for women's lingerie, the Thompsons decided to create a store that would cater to this need. As a response to the "ugly, floral-print nylon nightgowns" that were commonly found in department stores, they designed Victoria's Secret to provide a comfortable and luxurious shopping experience for women, offering a wide range of lingerie, womenswear, and beauty products. Today, Victoria's Secret is the largest American retailer of women's lingerie.

Table 17: Side-by-side example of the entity-perturbation (Longpre et al., 2021) and MACNOISE document samples. Text spans in green indicate the original answer text span, red is the entity-perturbed answer text span, including other text spans replaced by the perturbing sampled entity, and blue is the LLM-perturbed answer-text span; we also mark other text spans in blue in the MACNOISE context if they are related to the perturbed answer.

Question
who is the actor that plays tobias on nci

Answer
Joseph Peter Spano

Original Document
title: Joe Spano context: Joe Spano Joseph Peter Spano (born July 7, 1946) is an American actor who was originally best known for his role as Lt. Henry Goldblume on "Hill Street Blues". Spano is now best known for his work on "NCIS" as FBI Special Agent Tobias C. Fornell. Joe Spano was a member of the San Francisco improv group The Wing, and in college debuted as Paris in a production of "Romeo and Juliet" in 1967. In 1968, he helped found the Berkeley Repertory Theatre, appearing in its first production, and stayed with the company for 10 years. He moved to Hollywood

Entity Replacement (Longpre et al., 2021)
title: Joe Spano context: Joe Spano Jaeden Lieberher (born July 7, 1946) is an American actor who was originally best known for his role as Lt. Henry Goldblume on "Hill Street Blues". Spano is now best known for his work on "NCIS" as FBI Special Agent Tobias C. Fornell. Joe Spano was a member of the San Francisco improv group The Wing, and in college debuted as Paris in a production of "Romeo and Juliet" in 1967. In 1968, he helped found the Berkeley Repertory Theatre, appearing in its first production, and stayed with the company for 10 years. He moved to Hollywood

MacNoise (Perturbation Type: Additional Context)
title: Michael Thomas Grant context: Michael Thomas Grant, born on July 7, 1946, is an American actor who gained initial recognition for his role as Lt Henry Goldblume on "Hill Street Blues". He is now best known for his portrayal of FBI Special Agent Tobias C Fornell on the popular television series "NCIS". Grantś acting journey began in San Francisco as a member of the improv group The Wing, and he made his college debut as Paris in a 1967 production of "Romeo and Juliet". In 1968, Grant co-founded the Berkeley Repertory Theatre, where he appeared in its first production and remained with the company for a decade. Eventually, he moved to Hollywood and expanded his acting career to include television and film Some of his notable appearances include "Law & Order: Special Victims Unit", "The Good Wife", and "Blue Bloods". Apart from acting, Grant is a talented singer and musician, often displaying his abilities in stage productions and live performances.

Question
what is the southern most part of canada

Answer
Middle Island

Original Document
title: Geography of Canada context: will likely cause significant changes to the environment, including loss of the polar bear, the exploration for resource then the extraction of these resources and an alternative transport route to the Panama Canal through the Northwest Passage. The northernmost point within the boundaries of Canada is Cape Columbia, Ellesmere Island, Nunavut . The northernmost point of the Canadian mainland is Zenith Point on Boothia Peninsula, Nunavut . The southernmost point is Middle Island, in Lake Erie, Ontario (41’41’N, 82’40’W); the southernmost water point lies just south of the island, on the Ontario–Ohio border (41’40’35″N). The southernmost point of the Canadian

Entity Replacement (Longpre et al., 2021)
title: Geography of Canada context: will likely cause significant changes to the environment, including loss of the polar bear, the exploration for resource then the extraction of these resources and an alternative transport route to the Panama Canal through the Northwest Passage. The northernmost point within the boundaries of Canada is Cape Columbia, Ellesmere Island, Nunavut . The northernmost point of the Canadian mainland is Zenith Point on Boothia Peninsula, Nunavut . The southernmost point is Mid Glamorgan, in Lake Erie, Ontario (41’41’N, 82’40’W); the southernmost water point lies just south of the island, on the Ontario–Ohio border (41’40’35″N). The southernmost point of the Canadian

MacNoise (Perturbation Type: Entity Replacement w/ Context Match)
title: Geography of Canada context: The environment of Canada is undergoing significant changes, including the potential loss of the polar bear, increased exploration and extraction of resources, and the development of an alternative transport route to the Panama Canal through the Northwest Passage. The northernmost point within Canada’s boundaries is Cape Columbia, Ellesmere Island, Nunavut, while the northernmost point of the Canadian mainland is Zenith Point on Boothia Peninsula, Nunavut. The southernmost point of Canada is South Point Island, situated in Lake Erie, Ontario (41’41’N, 82’40’W). The southernmost water point lies just south of the island, along the Ontario–Ohio border (41’40’35″N).

Table 18: Side-by-side example of the entity-perturbation (Longpre et al., 2021) and MACNOISE document samples. Text spans in green indicate the original answer text span, red is the entity-perturbed answer text span, including other text spans replaced by the perturbing sampled entity, and blue is the LLM-perturbed answer-text span; we also mark other text spans in blue in the MACNOISE context if they are related to the perturbed answer.