Synthetic Paths to Integral Truth: Mitigating Hallucinations Caused by Confirmation Bias with Synthetic Data

Changwon Ok[*]
KT Corporation,
Republic of Korea
ok.changwon@kt.com

Eunkyeong Lee[*]
KT Corporation,
Republic of Korea
ek.lee@kt.com

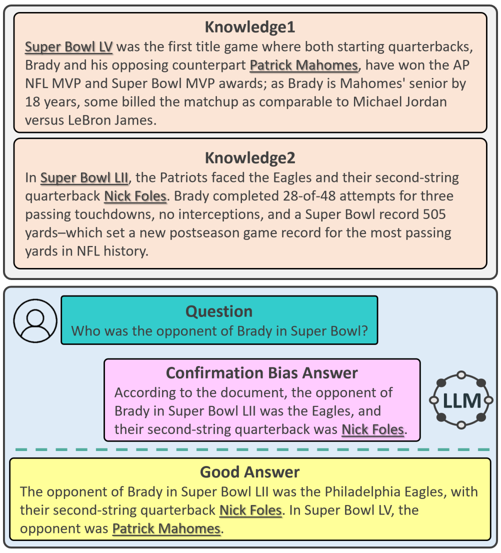
Dongsuk Oh[†]
Department of
English Language and Literature,
Kyungpook National University,
Republic of Korea
inow3555@knu.ac.kr

Abstract

Recently, large language models (LLMs) have made significant progress through retrieval-augmented generation (RAG) and preference learning. However, they still exhibit issues such as confirmation bias, the tendency to favor information that confirms one's beliefs, which remains largely unexplored in current research. In this paper, we propose a novel approach to mitigate confirmation bias-induced hallucination in LLMs through a synthetic data construction pipeline and Direct Preference Optimization (DPO) training. Our method enhances the integration of diverse and complementary information from multiple passages retrieved by RAG, enabling more balanced and accurate reasoning. Experimental results demonstrate significant improvements in response accuracy and reduced hallucination on benchmarks such as Natural Questions Open and HaluBench. These findings suggest that our approach effectively mitigates confirmation bias in long-context question answering, with potential applications to other NLP tasks. We release our data, and evaluation/train code for public access.[‡]

1 Introduction

Recently, large language models (LLMs) (Jiang et al., 2024a; Dubey et al., 2024; Minaee et al., 2024) have demonstrated remarkable success in various natural language processing (NLP) tasks, ranging from machine translation (Pourkamali and Sharifi, 2024) and summarization (Ravaut et al., 2024) to complex question answering and reasoning (Jiang et al., 2024b; Zhu et al., 2024). Despite these achievements, challenges persist, particularly when these models generate responses based on incomplete or ambiguous inputs (Tonmoy et al., 2024).


Knowledge1
Super Bowl LV was the first title game where both starting quarterbacks, Brady and his opposing counterpart Patrick Mahomes, have won the AP NFL MVP and Super Bowl MVP awards; as Brady is Mahomes' senior by 18 years, some billed the matchup as comparable to Michael Jordan versus LeBron James.

Knowledge2
In Super Bowl LII, the Patriots faced the Eagles and their second-string quarterback Nick Foles. Brady completed 28-of-48 attempts for three passing touchdowns, no interceptions, and a Super Bowl record 505 yards–which set a new postseason game record for the most passing yards in NFL history.

Question
Who was the opponent of Brady in Super Bowl?

Confirmation Bias Answer
According to the document, the opponent of Brady in Super Bowl LII was the Eagles, and their second-string quarterback was Nick Foles.

Good Answer
The opponent of Brady in Super Bowl LII was the Philadelphia Eagles, with their second-string quarterback Nick Foles. In Super Bowl LV, the opponent was Patrick Mahomes.

Figure 1: Example of Confirmation Bias in LLMs (Llama-3-8B-Inst) using RAG-retrieved Knowledge about *Brady's Super Bowl* appearances. The LLMs typically generate a biased answer by focusing only on one event, such as *Super Bowl LII*, while ignoring other relevant information like *Super Bowl LV*. The good answer represents the ideal response the model should generate, referencing both events to provide a complete and accurate answer.

Methods like retrieval-augmented generation (RAG) (Lewis et al., 2020a; Fan et al., 2024) have been developed to address these limitations. RAG enhances the accuracy of LLMs by incorporating external knowledge sources during the generation process. By retrieving relevant information from large databases or documents, RAG systems improve the model's ability to produce factually accurate outputs and mitigate hallucinations by grounding responses in external data. However, while RAG improves the factual accuracy of responses, LLMs may still suffer from confirmation bias, which leads them to generate biased responses by favoring specific retrieved information over others.

[*] These authors contributed equally.

[†] This author is corresponding author.

[‡] https://github.com/OccasionallyNLP/Synthetic-Paths-to-Integral-Truth.git

Confirmation bias (Nickerson, 1998) refers to the tendency to selectively process information that aligns with pre-existing beliefs or hypotheses while disregarding contradictory data. This cognitive bias, commonly observed in humans, can also manifest in artificial intelligence (AI) systems. When RAG retrieves a large quantity of information, the LLM may prioritize certain details that align with a specific narrative. While these details may be accurate within a narrow or limited context, they might not reflect a more comprehensive understanding of the information, potentially leading to biased responses that fail to account for the broader context.

For example, as shown in Figure 1, if a RAG model retrieves two pieces of information about Tom Brady's Super Bowl appearances and is asked the question, "*Who was the opponent of Brady in the Super Bowl?*," a biased model might respond, for example, "*The opponent of Brady in Super Bowl LII was the Eagles, and their second-string quarterback Nick Foles.*" while ignoring the first Super Bowl mentioned in the data. A more accurate response would be: "*The opponent of Brady in Super Bowl LII was the Philadelphia Eagles, with their second-string quarterback Nick Foles. In Super Bowl LV, the opponent was Patrick Mahomes.*" This example illustrates how confirmation bias can degrade the quality of a model's output, potentially leading to hallucinations by over-relying on specific retrieved information without considering the full context. To improve the accuracy and robustness of LLMs, it is essential to eliminate confirmation bias and ensure models can accurately utilize information from long contexts, even when complementary information is present.

This paper proposes a novel approach to help LLMs effectively address confirmation bias that arises in long, multi-passage contexts for question-answering tasks. To this end, we present a new strategy for constructing synthetic data that ensures more comprehensive and integrated reasoning, and through direct preference optimization (DPO) (Rafailov et al., 2024), we enhance the model's ability to generate results with reduced confirmation bias.

The contributions of this work are threefold:

- We propose a DPO training method utilizing synthetic data that mitigates confirmation bias, allowing LLMs to make accurate inferences even when exposed to complementary information in long contexts.

- We demonstrate that our method maintains performance across both long and short or single-passage scenarios, ensuring the model's ability to handle extensive information does not degrade.

- Finally, we provide a synthetic dataset, which can further improve LLM performance while reducing confirmation bias.

2 Related Works

2.1 Hallucination Mitigation with LLM

Retrieval Augmented Generation (RAG) is a technique that enhances the factuality of large language models (LLMs) by incorporating an external retrieval mechanism into the generative process (Tonmoy et al., 2024).

Lewis et al. (2020b) introduces RAG, demonstrating how integrating retrieval mechanisms into the generation process can improve factuality and reduce hallucinations. By leveraging external knowledge sources, models can produce more reliable and verifiable outputs, making them particularly effective for knowledge-intensive tasks. Glass et al. (2021) explore the application of RAG in zero-shot settings, where the model must perform tasks without task-specific training data. However, RAG is not without its limitations. The necessity to retrieve multiple passages to ensure comprehensive coverage can lead to increased prompt length, which in turn can degrade the model's performance due to context dilution and increased processing complexity. To address these limitations, prior research Tan et al. (2024) has explored methods that improve performance by integrating results generated by a generator with passages retrieved by a retriever. However, such approaches primarily focus on aligning the contexts of the generated content and the retrieved passages. Our study specifically addresses the issue of confirmation bias that occurs within the retrieved passages themselves.

2.2 Long Context with LLM

Although recent LLMs are capable of handling inputs up to 128K (Dubey et al., 2024), we have observed that they do not fully understand the passage when performing instructions based on long passages. For example, in a closed-QA task, response accuracy decreased depending on the location of the passage containing the correct answer. (Liu et al., 2024)

A solution to better understand long passages has emerged that utilizes LLM to augment the data based on long passages and supervised fine-tuning. Data augmentation methods have been proposed to augment long passages by generating queries and extracting responses based on tasks that can be performed in long passages (Bai et al., 2024), dividing long passages into segments and generating instruction-response pairs, and generating multi-hop QA pairs based on multiple segments (An et al., 2024). However, many data augmentations, such as 10K and 14K, are required for supervised fine-tuning. Therefore, we try preference learning to achieve effective alignment with less data.

2.3 Preference Learning with LLM

Once LLM could understand and perform a variety of difficult instructions requested by humans, attention turned to aligning it with human preferences to provide more useful, less harmful, and preferred responses. Proximal Policy Optimization algorithms (PPO) (Schulman et al., 2017), which are reinforcement learning, have been used for this purpose and have shown successful performance (Bai et al., 2022; Ouyang et al., 2022). Using reinforcement learning, complex and useful behaviors can be elicited, such as the ability to discriminate useful knowledge from a long input and answer it with a pre-trained weight (Zha et al., 2023).

PPO explicitly specifies the reward model and uses it to train the model. However, labeling reward data is difficult. There are three main types of data for reward models: point-wise, pair-wise, and ranking. DPO (Rafailov et al., 2024) is a method that allows for simple learning by computing the pairwise logit between selected and rejected pairs as a reward. In our paper, we use the basic DPO methodology.

2.4 Synthetic Dataset for Preference Learning with LLM

There is a many of research that utilizes LLM to generate synthetic preference data. The basic way to construct a preference dataset is to give LLM a generative sampling option to extract multiple responses, and then utilize a trained reward model to select the best/worst responses based on their scores to form selected and rejected pairs. There is a way to construct a chosen-rejected pair without a reward model by requesting a reward score in LLM-as-judge (Yuan et al., 2024). Instead of requesting a reward score, you can also build a pairwise dataset by presenting the LLM with a specific criterion, such as truthfulness, and asking it to choose the better of two answers (Tian et al., 2023). Another approach, similar to our paper, is to construct completions based on contrasting positive and negative prompts (Yang et al., 2023).

3 Proposed Method

Our hypothesis posits that an LLM generates higher-quality responses when it reflects and integrates all relevant segments of knowledge from the given context when answering questions. In contrast, responses that exhibit confirmation bias lead to diminished quality.

In this section, we propose a method for creating a dataset that captures this hypothesis without relying on human annotation for question answering, which is then followed by DPO (Rafailov et al., 2024). See Figure 2 for an overview of our proposed method.

In §3.1, we describe the key properties of the dataset, ensuring they align with the overall hypothesis of this research. Following this, in §3.2, we introduce the dataset creation pipeline, which automatically generates datasets from a provided corpus. This section includes the processes of Chosen Response Generation, Rejected Response Generation, and application to a real world question answering dataset. Finally, in §3.3, we present the DPO method, demonstrating how the constructed dataset is used for model fine-tuning and performance optimization.

3.1 Dataset for Mitigating Confirmation Bias

Confirmation bias is the tendency to favor evidence that supports existing beliefs or expectations (Nickerson, 1998). We argue that confirmation bias arises in two specific forms within our task:

- **Partial Evidence-Based Responses** These occur when the model generates responses using only a subset of the knowledge segments from the provided context.

- **Distorted Evidence-Based Responses** These occur when the model produces responses that contradict or misinterpret the provided context.

Both types can lead to hallucinations, where the generated content is nonsensical or unfaithful to the original source material (Ji et al., 2023).

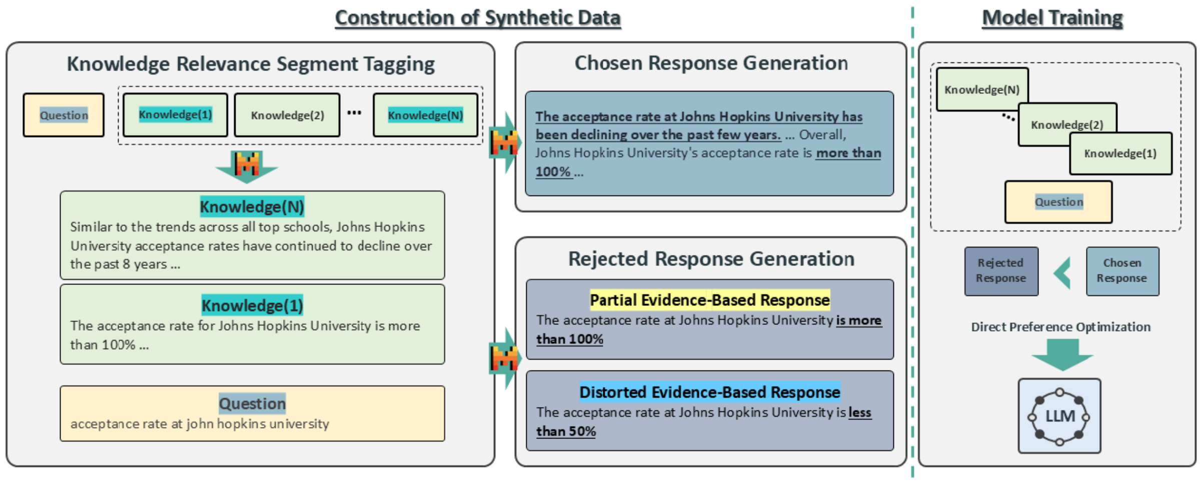

Construction of Synthetic Data
Model Training
Knowledge Relevance Segment Tagging
Question
Knowledge(1)
Knowledge(2)
...
Knowledge(N)
Knowledge(N)
Similar to the trends across all top schools, Johns Hopkins University acceptance rates have continued to decline over the past 8 years ...
Knowledge(1)
The acceptance rate for Johns Hopkins University is more than 100% ...
Question
acceptance rate at john hopkins university
Chosen Response Generation
The acceptance rate at Johns Hopkins University has been declining over the past few years. ... Overall, Johns Hopkins University's acceptance rate is more than 100% ...
Rejected Response Generation
Partial Evidence-Based Response
The acceptance rate at Johns Hopkins University is more than 100%
Distorted Evidence-Based Response
The acceptance rate at Johns Hopkins University is less than 50%
Knowledge(N)
Knowledge(2)
Knowledge(1)
Question
Rejected Response
Chosen Response
Direct Preference Optimization
LLM

Figure 2: **Synthetic Paths to Integral Truth.** The process of constructing synthetic data to mitigate confirmation bias has a total of three steps. (i) Knowledge Segment Relevance Tagging : Tagging the knowledge related to the question using LLM. The knowledge highlighted in blue green represents the knowledge segments related to the question. (ii) Chosen Response Generation : Generating a chosen response using only the knowledge related to the question (iii) Rejected Response Generation : Partial Evidence-Based Response, Distorted Evidence-Based Response using two methods to make rejected responses. The pair dataset is then used to optimize the LLM's preference.

We hypothesize that responses that reflect and integrate all relevant knowledge segments (referred to as chosen responses) are of higher quality, while those exhibiting confirmation bias (refered to as rejected responses) lead to low quality outputs.

3.2 Generating Synthetic Dataset

The dataset is generated automatically, following a process designed to meet the criteria outlined in §3.1. This process is organized into a three-stage pipeline: (1) Knowledge Segment Relevance Tagging, (2) Chosen Response Generation, and (3) Rejected Response Generation.

To process each step, we utilize prompt engineering with LLM, and further details used in this process are provided in Table 5. We use Mixtral-8x7B-Instruct-v0.1 for this process (Jiang et al., 2024a).

Knowledge Segment Relevance Tagging Our task is question answering using a given context. We first need to identify question-related knowledge segments from the given context to generate chosen/rejected responses. Since human annotation is time-consuming and expensive, we use an LLM to automatically annotate the question-related knowledge segments. For example, as shown in Figure 2, knowledge relevance tagging involves identifying and tagging the two knowledge segments in the given context that are relevant to the question, "acceptance rate at Johns Hopkins University." These knowledge segments contain information that can be used to infer the correct answer to the question. Specifically, "Knowledge (N)" mentions the trend of declining acceptance rates at Johns Hopkins University over the past 8 years, offering contextual information about the university's competitive admissions process. Meanwhile, "Knowledge (1)" directly addresses the acceptance rate of "more than 100%".

Chosen Response Generation A chosen response reflects and integrates all relevant knowledge segments in the given context. We select the question-related knowledge segments in the given context and prompt an LLM with the knowledge segments and the question itself to generate these chosen responses. As shown in Figure 2, the chosen response generation considers both question-related knowledge segments, "Knowledge (1)" and "Knowledge (N)" to generate the final answer, reflecting the information that the acceptance rate is "more than 100%" and has been "declining over the past few years.".

Rejected Response Generation To create rejected responses, we generate two types of outputs. The first type, **Partial Evidence-Based Response** is a response generated by providing the LLM with a prompt containing only a single knowledge segment related to the question, along with the question itself. The second type, **Distorted Evidence-Based Response**, introduces bias into the model's generation process, prompting it to generate an incorrect answer based on a single, manipulated knowledge segment. In Figure 2, the **Partial Evidence-Based Response** is generated by considering only the information from "Knowledge (1)", specifically "The acceptance rate for Johns Hopkins University is more than 100%." In contrast, the **Distorted Evidence-Based response** is created by incorrectly using the information from "Knowledge (1)" and altering it to "less than 50%".

Applying Pipeline to Question Answering Dataset To construct a dataset that reflects our hypothesis, we utilized the MS MARCO dataset [*](Nguyen et al., 2016). We use randomly sampled 1k instances in MS MARCO. We utilize a BM25 retriever to search for passages relevant to the given query, because the context from the original MS MARCO dataset is short [†]. We refered to the collection of the original context and the retrieved passages as a single context, and each part is considered a knowledge segment. To make the various data lengths, the number of retrieved passages is adjusted so that the prompt fits within the specified token length. Each token length is uniformly sampled from the set 1k, 2k, 4k, 8k. Additionally, to prevent the model from focusing on knowledge segments in a specific order, the sequence of knowledge segments within the context is randomized.

3.3 DPO with Synthetic Paths to Truth

We use the most popular alignment method, direct preference optimization (DPO). This method presents a second approximation that enables policy learning using only the chosen-rejected pairwise dataset instead of the reward model. Synthetic dataset pair set $\mathcal{D}$ for chosen and rejected pair-wise generation consists of (x, y_w, y_l) Here, x means prompt, y_l means rejected(losing) response, y_w means chosen(winning) response.
Given question, and knowledge segments $K = \{k_1, \cdots, k_N\}$ where N represents the number of knowledge segments related to question q. N can vary because the token length differs for each data sample. x is composed by concatenating the knowledge segments (K) and the question(q).
The loss of DPO incorporating our methodology can be expressed as follows:

$$\mathcal{L}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \\ [\sigma(\beta log\frac{\pi_\theta(y_w|q,k_1,\cdots,k_N)}{\pi_{ref}(y_w|q,k_1,\cdots,k_N)} \\ - \beta log\frac{\pi_\theta(y_l|q,k_1,\cdots,k_N)}{\pi_{ref}(y_l|q,k_1,\cdots,k_N)})] \quad (1)$$

here, σ is logistic function. β is a coefficient that controls the difference between the reference model and the policy model.
We randomly sample Partial Evidence-Based Responses and Distorted Evidence-Based Responses for $y_l \sim Unif\{y_l^{partial}, y_l^{distort}\}$ thereby constructing (x, y_w, y_l) triplet.

4 Experiment

4.1 Experimental Setups

We conduct our experiments using two main datasets: **NQ-Open** and **HaluBench**, along with several modified or derived subsets designed to investigate how the number and relevance of passages affect model hallucinations and confirmation bias.

4.1.1 Original Datasets

Natural Questions Open (NQ-Open) NQ-Open (Lee et al., 2019; Kwiatkowski et al., 2019) consists of real user queries submitted to the Google search engine, paired with human-annotated answers sourced from Wikipedia. Each data instance includes:

- **Gold passage**: A single passage containing the correct answer.
- **Distractor passages**: Passages that are relevant to the query but do not contain the correct answer.

We follow the setup described in Liu et al. (2024), where the number of retrieved passages (e.g., 10, 20, 30, 40) and their ordering (i.e., the position of the gold passage) can vary. All passages are retrieved using a fine-tuned retriever (Contriever fine-tuned on MS-Marco).

HaluBench HaluBench (Ravi et al., 2024) is an open-source benchmark comprising 15,000 (document, question, answer) triplets, originally designed for evaluating hallucination detection models. It emphasizes complex, real-world document-based question answering in domains such as fi-

[*] https://huggingface.co/datasets/microsoft/ms_marco
[†] https://github.com/castorini/pyserini

nance and medicine. Although HaluBench primarily targets hallucination detection, we adapt it to prompt models for answer generation. The model-generated answers are then evaluated for hallucination.

4.1.2 Derived Settings from NQ-Open

We create three derived datasets/settings from NQ-Open to isolate and analyze the effects of context length, passage count, and the presence of multiple relevant passages on model hallucination.

Lost in the Middle Setting In the original Liu et al. (2024) dataset, both the number of passages and the gold passage's position were varied extensively, resulting in a large number of data points. We adapt their original setting, observing how increasing the context size (number of passages) and altering the gold passage position may cause the model to overlook the correct passage, potentially leading to hallucinated answers.

Scaling with the Number of Passages Setting To investigate the effect of increasing the number of passages on hallucination, we modify the original dataset as follows:

- We randomize the gold passage position to focus solely on the effect of passage count rather than fixed ordering.
- We increase the context size up to 40 passages.
- To examine confirmation bias, we tag each passage using Mixtral-8x7B-Instruct-v0.1 to identify passages (other than the gold one) that could provide the answer.

This yields a dataset containing not only the gold passage but also multiple passages relevant to the query. Thus, we can analyze how scaling the total number of passages influences hallucination patterns.

Scaling with the Number of Relevant Passages Setting Using the 40-passage dataset described above, we further refine the data to investigate how multiple relevant passages affect confirmation bias:

- **Single**: Only one passage is relevant.
- **Multiple**: Two or more passages are relevant.

For each category, we sample 500 examples. Unlike the previous setting, the model here only receives the relevant passages at inference time. This approach allows us to understand:

- In the **Single** scenario, whether the model distorts information even with a single relevant passage.
- In the **Multiple** scenario, whether the model selectively uses only some of the relevant evidence, demonstrating confirmation bias.

4.1.3 Model and Training Details

We use the **Llama-3-8B-Instruct**[‡], trained with a context length of 8,192 tokens, as our baseline. We apply direct preference optimization (DPO) fine-tuning on this model using our constructed datasets (details are provided in the Appendix).

4.1.4 Evaluation Metrics

We employ the following metrics to assess model performance, hallucinations, and confirmation bias:

Accuracy Following Liu et al. (2024), accuracy measures whether the generated answer includes any correct solution.

Knowledge F1 (KF1) Based on Shuster et al. (2021), KF1 measures unigram overlap between the generated answer and the gold knowledge segment.

LLM as Judge Following Zheng et al. (2023), we use an LLM (Mixtral-8x7B-Instruct-v0.1) to assess whether the generated response is grounded in the provided context. Instead of simple lexical overlap, this method attempts to mimic human evaluation. For fairness, the LLM sees only the query-related context segments.

Lynx Score For the HaluBench dataset, we utilize the Lynx model (Ravi et al., 2024), trained to detect hallucinations by verifying the answer's faithfulness to the given document and question. We refer to this metric as the Lynx score, using an 8B model. This score complements the LLM as Judge approach and provides a specialized, model-based hallucination detection measure for the HaluBench domain.

4.2 Results in Question Answering

Table 1 presents a comparison between our model and the baseline, Llama-3-8B-Instruct. Our model outperforms Llama-3-8B-Instruct across all passage counts in the Scaling with the Number of Passages. With a single passage, our model achieves 92.32 accuracy, significantly surpassing the baseline's 84.56. This advantage persists even with 40

[‡] https://huggingface.co/Meta-llama/Meta-Llama-3-8B-Instruct

Model	# of Passages	Accuracy	KF1	LLM as Judge
Llama-3-8B-Inst	1	84.56	27.13	93.22
	10	65.35	30.24	62.2
	20	63.05	30.42	56.55
	30	59.55	29.34	55.83
	40	58.68	28.62	51.18
Ours	1	**92.32**	**55.53**	**97.66**
	10	**74.12**	**47.02**	**69.00**
	20	**69.11**	**45.83**	**63.24**
	30	**66.67**	**44.54**	**59.65**
	40	**65.76**	**43.78**	**57.14**

Table 1: The effect of changing the number of passages in Scaling with the Number of Passages. The position of the gold knowledge segment is randomized. Our model shows better performance than baseline in terms of accuracy and hallucination across all numbers of passages. Bold text indicates superior performance in the same condition.

Model	Lynx Score
Llama-3-8B-Inst	83.10
Ours	**89.11**

Table 2: Experiment results of HaluBench. Bold text indicates superior performance in the same condition.

passages, where our model maintains 65.76 accuracy versus the baseline's 58.68. In terms of mitigating hallucinations, our models' KF1 scores indicate a better overlap with the gold knowledge segment across all passage counts than the baseline.

In LLM as Judge score, the result also shows our model produces more aligned and grounded answers to the given context. Notably, The score given by the LLM as a judge shows a larger difference between the baseline and our model when the number of passages is multiple compared to when there is only a single passage, indicating that the difference increases as the number of passages shifts from single to multiple.

These results demonstrate that our fine-tuning approach effectively reduces confirmation bias and enhances factual consistency, making our model more robust in long-context question-answering tasks.

Table 2 shows that our model achieves a Lynx Score of 89.11, surpassing the baseline Llama-3-8B-Instruct, which scores 83.10. This improvement indicates our model's enhanced ability to generate responses that align more accurately with the source documents, effectively mitigating hallucinations. The superior performance can be attributed to our fine-tuning approach using DPO, enabling the model to integrate factual information more reliably. These results highlight the success of our model in reducing hallucinations, especially in complex, real-world scenarios like those present in the HaluBench dataset.

4.3 Lost in the Middle

Figure 3 measures the accuracy as a function of the position of the answer in the Lost in the Middle setting. If this effect were absent, the graph would appear as a flat, constant function. As shown in Figure 3, our model demonstrates a more gradual slope than Llama-3-8B-Inst and consistently achieves higher accuracy at every position.

When viewed in relative terms, both Llama-3-8B-Inst and our model have their highest accuracy at the 0th index (the first knowledge segment in the knowledge segments is the gold passage). However, compared to the accuracy at 0th index, our model experiences a smaller decrease in accuracy across all other indices than Llama-3-8B-Inst does. In example, When the maximum number of passages is set to 10, the accuracy for 4th index in Llama-3-8B-Inst is 89.61 and for our model is 94.46. And the accuracy for 9th index in Llama-3-8B-Inst is 81.71 and for our model is 88.83.

This result shows that our model has robust performance even if the number of passages is increased and the gold passage position is altered.

4.4 Diving into Mitigating Hallucination in Terms of Confirmation Bias.

Table 3 highlights that our model outperforms the baseline (Llama-3-8B-Inst) across all metrics in Scaling with the Number of Relevant Passages. In

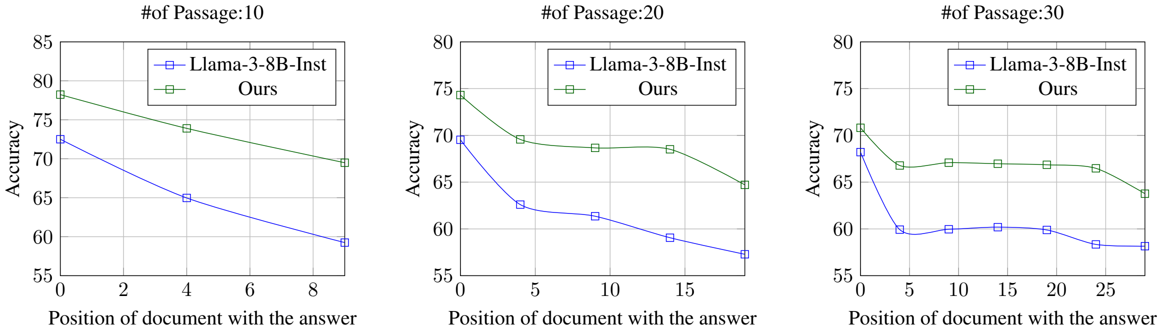


Figure 3: Comparison between our model and the Llama-8B-Inst baseline as the position of the document containing the answer increases in Lost in the Middle setting. The index on the x-axis represents the position of the gold passage among the total knowledge segments, and the index starts at 0.

Model	Type	Accuracy	KF1	LLM as Judge
Llama-3-8B-Inst	**Single**	86.80	28.04	90.00
	Multiple	63.60	32.49	48.71
Ours	**Single**	**94.20**	**54.52**	**96.20**
	Multiple	**73.00**	**55.99**	**58.56**

Table 3: Experiment results on the Scaling with the number of relevant passages, showing model performance based on the number of query-related passages. **Single** refers to instances with only one related passage, while **Multiple** indicates instances with multiple related passages. The results compare accuracy, KF1 score, and LLM-as-Judge evaluations between the Llama-3-8B-Inst and our model. Bold text indicates superior performance in the same condition.

the **Single** category, where only one passage is relevant, our model achieves an accuracy of 94.20 and a KF1 score of 54.52, significantly higher than the baseline's 86.80 accuracy and 28.04 KF1. For the **Multiple** category, which involves synthesizing information from multiple passages, our model achieves 73.00 accuracy and a KF1 of 55.99, compared to the baseline's 63.60 accuracy and 32.49 KF1. These results suggest that, even though we trained the model to mitigate confirmation bias, especially to avoid using partial evidence in the given context, it primarily learned to effectively utilize the correct knowledge segment when present.

In LLM as Judge score in **Single** category, the score improves from 90.0 to 96.20 (+Δ6.2) while the score improves from 48.71 to 58.56 (+Δ9.85) in **Multiple**. These results imply that our model effectively integrates information from all relevant sources, mitigating partial evidence-based responses. Furthermore, this suggests that our methodology can achieve a stronger effect in long-context scenarios.

Overall, these results show that our model successfully reduces hallucination related to confirmation bias, both in straightforward and more complex scenarios, by incorporating comprehensive evidence during response generation. We show model outputs in Appendix.

4.5 Assessing the Quality of Synthetic Data

Model	Accuracy	KF1
Llama-3-8B-Inst	58.68	28.62
Llama-3-8B-Inst (chosen)	62.64	40.40
Llama-3-8B-Inst (rejected)	39.59	27.86
Ours	**65.76**	**43.78**

Table 4: Experiment results in assessing the quality of synthetic data. Llama-3-8B-Inst (chosen) refers to the Llama-3-8B Inst model fine-tuned exclusively on the chosen data, while Llama-3-8B-Inst (rejected) refers to the same model fine-tuned exclusively on the rejected data. Bold text indicates the best performance, while underlined text indicates the second-best performance.

We did not directly evaluate the synthetic data we generated. Instead, we assessed its effectiveness indirectly by fine-tuning the Llama-3-8B-Inst model using synthetic data. Specifically, we trained models using only the chosen data, only the rejected data, and compared these with the original

Llama-3-8B-Inst model and our proposed model. The evaluation was conducted under the "Scaling with the Number of Relevant Passages" setting, focusing on scenarios with 40 passages.

Table 4 presents the evaluation results of various models regarding accuracy and KF1. The Llama-3-8B-Inst model serves as the baseline, while Llama-3-8B-Inst (chosen) and Llama-3-8B-Inst (rejected) are fine-tuned variants trained on the chosen and rejected synthetic data, respectively. Our model demonstrates the highest performance, achieving 65.76 in accuracy and 43.78 in KF1, surpassing all other models.

The improved accuracy and KF1 scores of Llama-3-8B-Inst (chosen), compared to Llama-3-8B-Inst, indicate that the chosen synthetic data enhances accuracy and reduces confirmation bias. In contrast, the lower accuracy and KF1 scores of Llama-3-8B-Inst (rejected) suggest that the rejected data increases confirmation bias and contains less reliable answers. These results highlight that the synthetic data we generated exhibit high quality in both its chosen and rejected subsets, supporting our approach to reducing confirmation bias.

5 Conclusion

In this work, we identified confirmation bias as a key factor contributing to hallucination in question answering under long-context scenarios. To address this, we proposed a dataset construction pipeline aimed at mitigating confirmation bias in both the chosen and rejected responses. Our method ensures the integration and effective utilization of knowledge segments that are closely related to the given question within the context. We then trained our model using the DPO-based approach.

In this paper, we have demonstrated the occurrence of confirmation bias in language models and proposed methodologies to mitigate it. For future work, it will be essential to investigate the underlying causes of confirmation bias in language models and provide empirical evidence to substantiate them. Additionally, we plan to explore the presence of confirmation bias in other tasks beyond question answering and investigate the application of our proposed methodology to these scenarios.

6 Limitations

The proposed method in the paper generates a DPO dataset for cases with gold knowledge. However, this is generated based on the assumption that the question is unconditionally answerable, so it can cause hallucinations in the unanswerable case. Depending on the performance of the RAG, only irrelevant knowledge may be retrieved, and generating data that accounts for this can further improve the model's ability to comprehend the retrieved knowledge. In addition, you can also consider the case where the knowledge contains contradictions.

Our proposed synthetic data distribution optimization methodology for mitigating confirmation bias in LLM models can be applied to an infinite number of tasks. We have verified its effectiveness in the multi-document question answering task, which is the easiest task to verify. We leave the experimentation of applying our method to various tasks such as summarization, document writing, editing, and rewriting as future work.

While the proposed method in the paper proved to be effective in mitigating confirmation bias, some issues may persist due to inherent vulnerabilities that naturally exist in humans.

Acknowledgments

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2022S1A5A2A03052880).

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *arXiv preprint arXiv:2404.16811*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. 2024. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on*

Knowledge Discovery and Data Mining, pages 6491–6501.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, and Alfio Gliozzo. 2021. Robust retrieval augmented generation for zero-shot slot filling. *arXiv preprint arXiv:2108.13934*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Feihu Jiang, Chuan Qin, Kaichun Yao, Chuyu Fang, Fuzhen Zhuang, Hengshu Zhu, and Hui Xiong. 2024b. Enhancing question answering for enterprise knowledge bases using large language models. *arXiv preprint arXiv:2404.08695*.

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.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *Preprint*, arXiv:1906.00300.

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

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset.

Raymond Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175–220.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Nooshin Pourkamali and Shler Ebrahim Sharifi. 2024. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *arXiv preprint arXiv:2401.08429*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Mathieu Ravaut, Aixin Sun, Nancy Chen, and Shafiq Joty. 2024. On context utilization in summarization with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2764–2781.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *Preprint*, arXiv:2407.08488.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hexiang Tan, Fei Sun, Wanli Yang, Yuanzhuo Wang, Qi Cao, and Xueqi Cheng. 2024. Blinded by generated contexts: How language models merge generated and retrieved contexts for open-domain qa? *arXiv preprint arXiv:2401.11911*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Finetuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023. Rlcd: Reinforcement learning from contrast distillation for language model alignment. *arXiv preprint arXiv:2307.12950*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Andrew Zhu, Alyssa Hwang, Liam Dugan, and Chris Callison-Burch. 2024. Fanoutqa: A multi-hop, multi-document question answering benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 18–37.

A Prompt Engineering for Dataset Generation

Table 5 shows our prompts used in our dataset generation pipeline.

Knowledge Segment Relevance Tagging In our dataset (selected MSMARCO), each dataset has the context which is composed of multiple knowledge segments (the original context and retrieved passages). To increase accuracy on the knowledge segment relevance tagging, we provide the large language model (Mixtral) with a single knowledge segment sequentially, rather than supplying the entire context at once for tagging.

Chosen response generation We use only the question-related knowledge segments that can be obtained from **Knowledge Segment Relevance Tagging**. As shown in Table 5, we concatenate the question-related knowledge segments as context and give LLM question and context to generate chosen response.

Rejected response generation We have two types of rejected response generation. First, **Partial Evidence-Based Response** is a response generated by only a single knowledge segment related to the question, along with the question. To make this response, we recycle the chosen response generation prompt. Unlike chosen response generation, we just make a single sampled question-related segment and question as prompt to make rejected response generation. The second type, a **Distorted Evidence-Based Response** generates an incorrect answer based on a single, manipulated knowledge segment. To do this, we use the prompt as shown in Table 5.

B LLM as Judge for hallucination

We use the LLM as a judge for hallucination, specifically to assess whether the generated response is grounded in the given context. We use the prompt as shown in table 6. Since the LLM might struggle to evaluate a given claim within a single long document, we sequentially feed each passage/knowledge segment feed into the prompt for the LLM to judge. When multiple judgments were obtained, we calculated the score for each data sample by dividing the number of ‘yes’ responses by the total number of judgments. Finally, the overall score for the entire dataset was obtained by averaging the individual LLM as judge scores.

Type	Prompt
Knowledge Segment Relevance Tagging	Please identify the given knowledge segment is related to the given question. Make answer just TRUE or FALSE. JUST answer the question. DO NOT say the explanation. Provide your response as follows: Answer: (TRUE or FALSE) Now here are the passage and question. Knowledge segment : {knowledge segment} Question: {question} Answer:
Chosen Responses	Answer the question based on the context. You SHOULD use all the information in the context to answer the question. SHOULD NOT say that you answered based on the given context. Context : {context} Question: {question} Answer:
Distorted Evidence-Based Responses	The Given context is UNRELATED to the question. Make WRONG answer to the question based on the UNRELATED context. SHOULD NOT say that you answered based on the given context. UNRELATED Context : {context} Question: {question} WRONG Answer:

Table 5: Prompts used in the dataset generation pipeline.

Prompt
Determine whether the provided claim is grounded in the corresponding document. In this context, grounding means that the information presented in the claim is substantiated by the document. If not, the claim should be considered not grounded. Document: document Claim: claim Please assess whether the claim is grounded in the document by responding with either ’yes’ or ’no.’ Answer:

Table 6: Prompts used in LLM as judge to evaluate groundness in the given claim (generated text) to the given document. To obtain more accurate results, we sequentially provided the model with only the passages related to the question from the given context as prompts.

Type	Question	Knowledge segment(s)	Llama-3-8B-Inst	Ours
Single	when do the walking dead comics come out	Title: The Walking Dead (comic book) The Walking Dead debuted in 2003, published by Image Comics, with art by Tony Moore for the first six issues and Cliff Rathburn shading the art after issue five.(. . .)	According to the document, there is no information about when the Walking Dead comics come out.	The Walking Dead comic book series debuted in 2003 and has been published by Image Comics.
Multiple	where do they film take me out fernando's	Title: Se Me Va la Voz to the script. Adriana Ontiveros, a model from Jalisco also participated in the video.(. . .) Title: Take Me Out (UK game show) (. . .) In the first series, the successful couples conducted their date at FERNANDO'S! in Manchester, which is actually Club Bijou on Chapel Street (. . .) For the second series, the date took place abroad on the Isle of FERNANDO'S!, a fictional location based in Puerto de la Cruz, Tenerife. (. . .)	According to the document, the answer is: Puerto de la Cruz, Tenerife.	For the Take Me Out Fernandos!, the date took place in two different locations. In the first series, the successful couples conducted their date at FERNANDO'S! in Manchester, which is actually Club Bijou on Chapel Street. In the second series, the date took place abroad on the Isle of FERNANDO'S!, a fictional location based in Puerto de la Cruz, Tenerife.

Table 7: Comparision in the performance of the Llama-3-8B-Inst model with our proposed model on both **Single** and **Multiple** knowledge segment(s) question answering tasks in 'Scaling with the number of relevant passages' dataset. The highlighted text in different colors indicates the specific parts of these segments that serve as evidence for the answer.

C Scaling with the number of relevant passages

Table 7 presents a comparison of responses between Llama-3-8B-Inst and our model for questions involving both single and multiple knowledge segments in 'Scaling with the number of relevant passages' dataset.

- **Single Question** The question is "When do The Walking Dead comics come out?". The relevant knowledge segment(red-highlighted text) describes that "The Walking Dead debuted in 2003". Llama-3-8B-Inst responds with, "There is no information about when The Walking Dead comics come out," failing to extract the relevant information and answer the question. In contrast, our model successfully identifies the relevant information from given context and answers properly, stating, "The Walking Dead comic book series debuted in 2003 and has been published by Image Comics."

- **Multiple Question:** The question is "Where do they film Take Me Out at Fernando's?" The knowledge segment provides two pieces of evidence: (1) In the first series, the filming location was "Club Bijou on Chapel Street," and (2) in the second series, it occurred "Puerto de la Cruz, Tenerife." Llama-3-8B-Inst gives an partial evidence based answer, "Puerto de la Cruz, Tenerife." However, our model captures the entire relevant knowledge segments and integrate them properly, stating both filming locations correctly.

This demonstrates the effectiveness of the our model in accurately extracting the information and integrating them to make answers, including specific details missed by Llama-3-8B-Inst.

D Training Details

Hardware Details The training of mitigating Hallucination model A100 80GB 1 node. The max length of the backbone model is 8192 tokens.

DPO hyperparameter Details we use Dubey et al. (2024)'s optimal hyperparameter. the optimal parameter was utilized with a learning rate of 1e-5, beta of 0.1 and global batch of 256.