CORD: Balancing COnsistency and Rank Distillation for Robust Retrieval-Augmented Generation

Youngwon Lee[*] **Seung-won Hwang**[*] **Daniel Campos**
Filip Graliński **Zhewei Yao** **Yuxiong He**
Snowflake AI Research [*]Seoul National University

Abstract

With the adoption of retrieval-augmented generation (RAG), large language models (LLMs) are expected to ground their generation to the retrieved contexts. Yet, this is hindered by position bias of LLMs, failing to evenly attend to all contexts. Previous work has addressed this by synthesizing contexts with perturbed positions of gold segment, creating a position-diversified train set. We extend this intuition to propose consistency regularization with augmentation and distillation. First, we augment each training instance with its position perturbation to encourage consistent predictions, regardless of ordering. We also distill behaviors of this pair, although it can be counterproductive in certain RAG scenarios where the given order from the retriever is crucial for generation quality. We thus propose CORD, balancing COnsistency and Rank Distillation: CORD adaptively samples noise-controlled perturbations from an interpolation space, ensuring both consistency and respect for the rank prior. Empirical results show this balance enables CORD to outperform consistently in diverse RAG benchmarks.

1 Introduction

Recently, large language models (LLMs) have incorporated retrievers to augment input contexts for more grounded generation. However, during retrieval-augmented generation (RAG), LLMs reportedly suffer from position bias where they pay disproportionate attention to different parts, worsened as the input becomes longer (Liu et al., 2024). An existing solution has synthesized a training set by randomizing the position of gold segment (An et al., 2024). It allows LLMs to implicitly learn that relevant information can appear at any position, mitigating position bias.

Our distinction is to pursue dual goals of (1) **CO**nsistency for mitigating position bias and (2) **R**ank **D**istillation, learning to utilize meaningful signals in the given order from the retriever and also to denoise it, for robust RAG.

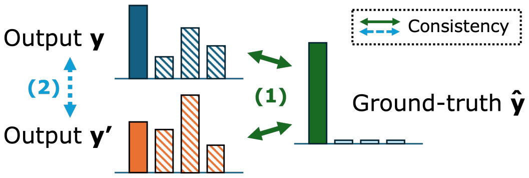

Figure 1: Enforcing consistency with (1) augmentation (green) and (2) distillation (blue).

Method	(A)	(B)
Given order	41.34	56.52
+ consistency	36.87 (↓)	57.87 (↑)
CORD (ours)	**44.74** (↑)	**58.71** (↑)

Table 1: Generation quality with different methods in representative RAG scenarios A and B, where distillation may hinder or enhance, respectively.

For CO, we extend the position-perturbing training intuition, by augmenting the retriever-given order of contexts $\mathbf{c}$ with its perturbation $\mathbf{c}'$, sharing the same ground truth $\hat{y}$. Green arrows in Figure 1 visualize how this augmentation indirectly enforces consistency by guiding predictions y from $\mathbf{c}$ and y' from $\mathbf{c}'$, to converge to the ground-truth $\hat{y}$.

To further enforce consistency, a distillation loss can be added to directly penalize the distributional divergence in all outputs. The blue arrow in Figure 1 visualizes this loss further incentivizing consistent internal representation, by distilling 'dark knowledge' (Hinton et al., 2015; Sadowski et al., 2015; Furlanello et al., 2018) from one to another.

However, pursuing CO objective alone, without balancing it with the RD objective, is counterproductive in some scenarios as illustrated in Table 1. It contrasts two representative real-life RAG scenarios A and B:[1] In A, retriever provides a reliable rank prior, such that distilling predictions from a

[*]Work done while visiting Snowflake. Correspondence to: seungwonh@snu.ac.kr.

[1]For presentation brevity, we reveal in later section.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 787–796
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

randomized ordering can unlearn this helpful prior, as evidenced by the degradation in generation quality after consistency regularization. Meanwhile, in B, where generation is not sensitive to the given order, CO objective enhances performance.

Our technical contribution is to adapt $\mathbf{c}'$ to the given scenario, by controlling the degree of perturbation, in place of $\mathbf{c}'$ with a fixed randomization. We define an interpolated space of perturbations and dynamically sample an appropriate level of perturbation from it. Table 1 shows CORD outperforms in both scenarios, by sampling smaller perturbations in scenario A, where rank prior is important, and larger perturbations in scenario B, where robustness to position bias is crucial.

Our contribution can be summarized as follows: (1) We propose CORD, balancing consistency and rank distillation in RAG. (2) We show distilling with a controlled perturbation, sampled from an interpolated space of teachers, is effective across 5 diverse RAG scenarios, whereas existing consistency methods fall short.

2 Related Work

2.1 Position Bias in Long Context LLMs

Liu et al. (2024) and similar works have shown that LLMs favor input contexts placed at the beginning or end of the input, a tendency that benchmarks such as needle-in-a-haystack[2] aim to assess by testing their ability to locate relevant information (*needle*) within long, potentially irrelevant contexts (*haystack*). An et al. (2024) extended this understanding by training models on synthetic data, intentionally perturbing a position of gold segment and adding random noises. Similarly, Fu et al. (2024) examined continual pretraining of LLMs on long-context data to expand their context window size for retrieving information.

Our distinction is to use position perturbation for a different objective of data augmentation for consistency training.

2.2 Data Augmentation for Consistency

Pairing a datapoint with a counterfactual applying a small perturbation has been mainly studied for robust training on simpler tasks such as classification (Xie et al., 2020). To our knowledge, we are the first to augment a position-perturbed retriever during training and enforce consistency for RAG.

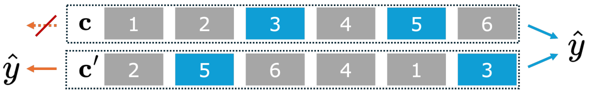

Figure 2: (Left) IN2 only uses $\mathbf{c}'$. (Right) We augment the given order $\mathbf{c}$ (top) with perturbed ranking $\mathbf{c}'$ (bottom) and use both.

Another related line of work is interpolating two training instances (Chuang and Mroueh, 2021), which we extend to define a space of controlled perturbations for dynamic adaptation in Section 3.2.

3 Method

3.1 CO: Consistency Regularization

We propose to mitigate position bias by regularizing output consistency over possible perturbations, through (1) augmentation and (2) distillation.

First, we explain how augmenting position-perturbed examples contributes to consistency. We first formalize RAG as generating an answer y given an input x,

$$y \sim p(\cdot \mid x, \mathbf{c}), \tag{1}$$

along with the sequence of n retrieved contexts $\mathbf{c} = [c_1; c_2; \cdots; c_n]$. Then, for a training triplet $(x, \mathbf{c}, \hat{y})$ the negative log-likelihood (NLL) loss for maximum likelihood estimation training is

$$\mathcal{L}_{\text{n}} = -\sum_t \log p(\hat{y}_t \mid x, \mathbf{c}, \hat{y}_{<t}), \tag{2}$$

which encourages the model to produce the correct answer $\hat{y}$ given the input x and retrieved contexts.

Inspired by An et al. (2024), referred to as IN2, we employ position perturbation to augment $\mathbf{c}$ from the corpus $\mathcal{C}$ with $\mathbf{c}'$. For comparison, IN2 synthesized question and context $(q, \mathbf{c})$ pairs where the gold passage s for generating the gold answer $\hat{y}$ appears in various positions. As Figure 2 shows, we retain both the original $(q, \mathbf{c}, \hat{y})$ and the perturbed examples $(q, \mathbf{c}', \hat{y})$: Unlike IN2's using $\mathbf{c}'$ only for training (orange arrows), we train over the augmented dataset $\mathcal{C}'$ which includes both $\mathbf{c}$ and $\mathbf{c}'$ (blue arrows). Predictions for both are supervised to converge to the same ground-truth $\hat{y}$ using the loss in Eq. 2.

Second, by adding a distillation loss, we can further match token-level output probability distributions for $\mathbf{c}$ and $\mathbf{c}'$. We use the sum of Jensen-Shannon Divergence (JSD) between output proba-

[2] github.com/gkamradt/LLMTest_NeedleInAHaystack

bility distributions at each time step t for this purpose:[3]

$$\mathcal{L}_{\text{d}} = \sum_{t} \text{JSD}\left(f_t(\mathbf{c}) \,\|\, f_t(\mathbf{c}')\right), \qquad (3)$$

where $f_t(\mathbf{c}) = p(\hat{y}_t \,|\, x, \mathbf{c}, \hat{y}_{<t})$. This encourages the model to align its internal representations of input and association with the output, encoded in the 'dark knowledge' (Hinton et al., 2015; Sadowski et al., 2015; Furlanello et al., 2018) across different perturbations.

Finally, the two types of loss in Eq. 2 and 3 can be combined to obtain our training objective:

$$\mathcal{L} = \mathcal{L}_{\text{n}} + \lambda \cdot \mathcal{L}_{\text{d}}, \qquad (4)$$

where the hyperparameter λ determines the relative strength of the two terms.

3.2 RD: Adaptive Teacher Selection for Rank Distillation

However, as previously outlined in Table 1(A), distill loss on a random perturbation $\mathbf{c}'$ may interfere with the RD objective in an RAG scenario where retriever provides a meaningful ranking $\mathbf{c}$ with valuable prior: In this work, we consider MS MARCO (Bajaj et al., 2018) as a representative example, where an industry-scale complex retrieval system provides the ranking.

Figure 3(A) depicts such unlearning of ranker prior, when distilled from a random perturbation in scenario A. The y-axis in the plot represents the probability the LLM assigns to the ground-truth answer, $p(\hat{y} \,|\, x, \mathbf{c})$ for the given order $\mathbf{c}$ (solid circle) and $p(\hat{y} \,|\, x, \mathbf{c}')$ for random perturbation $\mathbf{c}'$ (empty circle). In MS MARCO, the given order $\mathbf{c}$ carries a useful prior, resulting in high probability of the ground-truth $p(\hat{y} \,|\, x, \mathbf{c})$. Randomizing this order would greatly lower the probability $p(\hat{y} \,|\, x, \mathbf{c}')$, such that enforcing consistency between the two would unlearn the benefit of rank prior.

To tackle this, instead of fixing $\mathbf{c}'$ as a random perturbation, we define a sample space and strategy for adaptive teacher selection, to control the degree of perturbation for distillation. We introduce an interpolation of $\mathbf{c}$ and $\mathbf{c}'$ with a controlled noise degree of α, denoted as $\mathbf{c}'_{\alpha}$: Here, the lower ranked α proportion of the retrieved contexts is randomized while the remaining retains the given order. In Figure 3, such interpolated sample is shown as a shaded circle on a dotted line, the interpolated path connecting $\mathbf{c}$ and $\mathbf{c}'$, as the noise degree α varies from 0 to 1. For brevity, we assume a desirable single value of α for the given task is known a priori, and later discuss how to find it in Section 3.3.

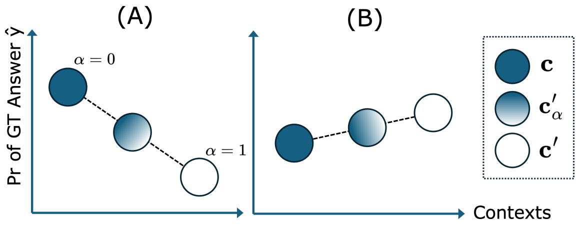

Figure 3: Interpolated sample space for scenario A and B from Table 1, where (A, left) perturbation leads to a large drop in probability of ground-truth $\hat{y}$, and (B, right) with no such drop.

This interpolation allows to select a better teacher between $\mathbf{c}'_{\alpha}$ and $\mathbf{c}'$ by choosing the one with a higher probability of predicting the ground truth. As shown in Figure 3A, small perturbations tend to yield higher y values in scenario A as they retain the given order in part, leading to $\mathbf{c}'_{\alpha}$ chosen for distillation. This corresponds to ensembling two retrievers, which agree on top-ranked documents but diversify the ranks of the rest.

An added advantage is, the same approach seamlessly supports scenario B, where there is no conflict between CO and RD. As illustrated in Figure 3(B), the y-axis score remains relatively stable across different orderings, and moreover, the score is no longer sensitive to ordering. Thus, pairing the given order with the one that has a higher y score essentially serves the goal of pursuing CO.

3.3 Score-Aware Teacher Sampling

So far, we have mainly focused on utilizing *rank* prior from the retriever; however, the retriever may provide varying level of information in different RAG scenarios, such as score for each item as well. We describe how to incorporate such additional signals into adaptive teacher sampling.

When no prior knowledge of the distribution of the probability of ground-truth $p(\hat{y} \,|\, x, \mathbf{c}'_{\alpha})$ over the interpolated path is known, we follow the principle of maximum entropy (Jaynes, 1957) to assume uniform distribution. That is, we choose to sample $\alpha = 0.5$ from the interpolated space defined in Section 3.2, where α varies in the range of $(0, 1)$.

Alternatively, we utilize retriever scores as a

[3] While we default to summing all terms, the number of time steps t to aggregate in Eq 3 can be adjusted for efficiency, as detailed in Appendix B.

	MS MARCO		HotpotQA		NQ		MN	MN-IDK
Finetuning Objective	R-L	GPT-4	EM	GPT-4	Acc	GPT-4	F_1	Acc
No finetuning	41.34	51.94	42.86	66.50	52.18	62.46	56.52	54.82
$\mathcal{L}_{\text{nll}}$ on $\mathcal{C}'$	44.52	**57.28**	58.62	83.75	55.60	63.51	56.25	95.78
CORD	**44.74**	**57.28**	**63.55**	**85.72**	**58.55**	**63.72**	**58.71**	**98.83**

Table 2: RAG performance with Phi-3 3B as the generator and different finetuning strategies applied.

	MS MARCO	
Finetuning Method	R-L	GPT-4
No finetuning	41.34	51.94
$\mathcal{L}_{\text{nll}}$ on $\mathcal{C}$	41.81	51.94
$\mathcal{L}_{\text{nll}}$ on $\mathcal{C}'$	44.52	**57.28**
CORD	**44.74**	**57.28**

Table 3: Without augmentation (second row) there is a clear performance gap compared to models trained with consistency objectives (third and fourth row).

	MN	MN-IDK
Finetuning Method	F1	Acc
CORD	58.71	98.83
+ Adaptive α	59.16	98.83

Table 4: Effect of dynamically adjusting α based on retriever score.

proxy for the unknown distribution of $p(\hat{y} \mid x, \mathbf{c}'_\alpha)$, from which the optimal noise level α can be determined. Specifically, we aim to extract the most confident top-ranked contexts identified by the retriever, by preserving the contexts ranked above the largest discontinuity in scores and perturbing the rest. Given scores s_i for each retrieved context $c_i \in \mathbf{c}$, which are sorted in descending order of score, i.e., $s_1 > s_2 > \cdots > s_n$, we locate the adjacent pair of passages with the largest gap in retriever score $\hat{i} = \text{argmax}_i(s_i - s_{i+1})$ and perturb the passages ranked lower than $\hat{i}$. In other words, we choose $\alpha = 1 - \hat{i}/n$ for this example. Intuitively, this approximates finding the largest acceptable degree of noise that would still result in sufficiently high $p(\hat{y} \mid x, \mathbf{c}'_\alpha)$.

4 Results

We design evaluations to answer these research questions:

- (RQ1) Does CORD pursue dual goals of CO and RD effectively?
- (RQ2) Does CORD adaptively choose $(\mathbf{c}, \mathbf{c}')$ pair in different scenarios?
- (RQ3) How can the noise degree α for interpolation be tuned per task or example?

4.1 Experimental settings

We have evaluated our proposed method on several QA benchmarks: MS MARCO (Bajaj et al., 2018), HotpotQA (Yang et al., 2018), NaturalQuestions (Kwiatkowski et al. (2019); NQ) as reorganized by Liu et al. (2024). We further consider multi-needle (MN) dataset, which is built following An et al. (2024), as a scenario where irrelevant contexts are prevalent and retriever prior is not meaningful.[4]

For evaluation, we used widely reported metrics for each benchmark, namely ROUGE-L for MS MARCO, exact match (EM) for HotpotQA, and span-based exact match, or 'accuracy' for NQ. We also adopted the evaluation protocol from Yang et al. (2024) using GPT-4, allowing more flexibility in answers. For MN where answers typically contain a few sentences, we report sentence-level F_1, and for MN-IDK, an unanswerable split of MN, we report accuracy. Further details can be found in Appendix A.

4.2 Results

Bias mitigation and rank distillation Table 2 shows that our proposed method outperforms the baselines across all benchmarks, validating its effectiveness in pursuing dual goals of CO and RD.

In addition, Table 3 shows the importance of denoising through consistency in rank distillation. There is a clear performance gap between the model trained on the given order $\mathbf{c}$ without augmentation (second row), and those augmented (third and fourth) on MS MARCO. This suggests that even with a strong rank prior, consistency across slight perturbation positively contributes to RD, by mitigating potential bias from retriever or generator.

Adaptive pair selection CORD indeed selects the proper teacher for enforcing consistency, while

[4] This corresponds to scenario B in Table 1 and Figure 3.

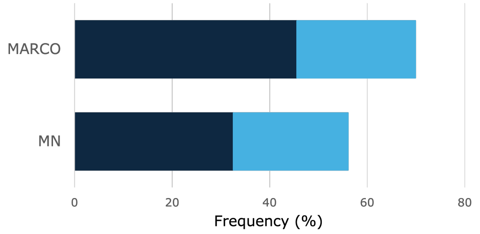

Figure 4: (Top) On MS MARCO, the interpolated noise-controlled perturbation $\mathbf{c}'_\alpha$ (dark blue) is much more likely to be paired with the given order $\mathbf{c}$, than $\mathbf{c}'$ (light blue). (Bottom) The gap is much smaller on MN.

the tendency in choices exhibit clear difference per different RAG scenario, as shown in Figure 4. The ratio of $\mathbf{c}'_\alpha$ paired with $\mathbf{c}$ is shown with dark blue, while the ratio of $\mathbf{c}'$ paired with $\mathbf{c}$ is presented by light blue bar. Comparing MS MARCO (top) and MN (bottom), it is clearly shown that $\mathbf{c}'_\alpha$ is much more likely to be paired with $\mathbf{c}$ in the former, where the RD objective is more prominent. This supports our rationale behind designing adaptive teacher selection in Section 3.2.

Score-aware teacher sampling Table 4 shows that score-aware dynamic adjustment of α, described in Section 3.3 brings further gain; the effective mean value of α throughout the train set was 0.8, suggesting a larger portion of the ranking was allowed to be perturbed.

5 Conclusion

We have presented CORD, to balance the tension between CO (consistency) and RD (rank distillation) objectives in RAG. For the former, we augment order-perturbed contexts and add distillation loss for explicit consistency regularization. For the latter, CORD adaptively chooses desirable degree of perturbation to prevent unlearning valuable prior from the retriever. CORD consistently outperforms existing methods in diverse RAG scenarios.

Limitations

Whether our findings generalize over diverse models can be further explored. In addition, the pros and cons of diverse mixing strategies for an interpolated sample space, such as employing another retriever for mix, can be explored; we leave it as future work.

References

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

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *Preprint*, arXiv:1611.09268.

Ching-Yao Chuang and Youssef Mroueh. 2021. Fair mixup: Fairness via interpolation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. 2024. Data engineering for scaling language models to 128k context. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1607–1616. PMLR.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

E. T. Jaynes. 1957. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630.

Rohan Jha, Bo Wang, Michael Günther, Saba Sturua, Mohammad Kalim Akram, and Han Xiao. 2024. Jina-colbert-v2: A general-purpose multilingual late interaction retriever. *CoRR*, abs/2408.16672.

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.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2024. The unlocking spell on base llms: Rethinking alignment via in-context learning. In *International Conference on Learning Representations*.

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.

Peter Sadowski, Julian Collado, Daniel Whiteson, and Pierre Baldi. 2015. Deep learning, dark knowledge, and dark matter. In *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning*, volume 42 of *Proceedings of Machine Learning Research*, pages 81–87, Montreal, Canada. PMLR.

Qizhe Xie, Zihang Dai, Eduard H. Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, Lingkun Kong, Brian Moran, Jiaqi Wang, Yifan Ethan Xu, An Yan, Chenyu Yang, Eting Yuan, Hanwen Zha, Nan Tang, Lei Chen, Nicolas Scheffer, Yue Liu, Nirav Shah, Rakesh Wanga, Anuj Kumar, Wen-tau Yih, and Xin Luna Dong. 2024. CRAG - comprehensive RAG benchmark. *CoRR*, abs/2406.04744.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Implementation Details

MN construction For MN data construction, we generally followed the recipe from An et al. (2024), with the subtle difference that Mixtral was used for question and answer generation. When preparing the MN dataset following An et al. (2024), we generally abide by their practices, while using Mixtral as the LLM for question and answer extraction, and employed GPT-4 to verify it. For the seed corpus, we utilized the same `realnewslike` subset from the C4 corpus as $\mathcal{C}$. We refer the reader to their original paper for more details.

In addition, to study how LLMs can be trained to refuse to answer when there are insufficient evidence provided, rather than to hallucinate, we split the test set into two settings, answerable and unanswerable: In the latter, dubbed MN-IDK, the gold segment s that provides the evidence to answer the given question is omitted. Thus, the model is expected to answer it does not have enough evidence in the contexts to provide the correct answer, or, 'I don't know.'

Metrics The evaluation protocol involving GPT-4 as the judge from Yang et al. (2024) evaluates the correctness of the answer with greater flexibility, compared to the canonical lexical match based metrics, and is known to align better with human judgment. Also, it penalizes hallucinated response more than simply abstaining.

While other benchmarks considered in this work require shorter answers, expected answers in MN and MN-IDK typically comprise of a few sentences: thus, we report sentence-level F1 score for MN, where GPT-4 was used as a judge in the same manner as the method described above to decide each sentence in the generated answer is supported by the ground-truth (precision), and vice versa (recall). For MN-IDK, GPT-4 determined whether the model response successfully refused to provide the answer or not, and we reported the accuracy.

Prompts provided to LLM for both type of evaluation can be found in Appendix C.

Training For MS MARCO, HotpotQA and MN, we finetuned Phi-3 3B model on their respective train data: for MS MARCO, we used 20k examples held out from v2.1 dev set for training, and used non-overlapping subset for testing.

For training with CORD on MN, as described in Section 3.2, we generated an artificial ranking over the passages by reranking them with a ColBERT variant model from Jina AI,[5][6] which also provided scores for each passage. This artificial ranking serves as the opposite extreme of the interpolated perturbation space, $\mathbf{c}'$.

The base model, Phi-3 3B, was trained with LoRA at bf16 precision. The relevant hyperparameter configuration was as follows: for LoRA related settings, we used rank of $r = 8$, $\alpha = 32$, and dropout rate of 0.1. For general configuration, we used linear decay for scheduling with initial learning rate of 1e-4 and effective batch size of 4; we trained the model for 5 epochs with weight decay of 0.01 applied. For CORD-specific configuration, we set coefficient for consistency loss strength λ as 10 and the noise degree for interpolating contexts α as 0.5 throughout our experiments. We leave it as future efforts to search for optimal configuration for these values per different scenarios.

B Design of Consistency Loss

Using the loss from the first token of the answer only also worked reasonably. We attribute this to that contribution of the consistency loss terms from earlier time steps, i.e., those from the beginning of the ground-truth, are larger than that of those from later time steps. The model output probability distribution for time step t defined previously in Section 3.1 is indeed conditioned on the shared prefix of the ground-truth answer $y_{<t}$: as more tokens in the prefix are conditioned in both sides as t increases, the distribution over the token to be immediately followed f_t would converge, as less and less options would be part of a plausible continuation of the answer. This results in terms from later t contributing smaller to the total loss $\mathcal{L}_{\text{con}}$, which is why dropping all of them but some at the beginning, just one in the extreme case, suffices to regularize the model output. It is consistent with the findings from previous papers showed that token-level distributional shift between the base and finetuned LLM decreases over time step during decoding (Lin et al., 2024).

While the benchmarks we have considered generally require rather short responses, it remains to see if this mechanism of using the first time step only for consistency loss computation also work well for long-form answer generation tasks.

[5] `huggingface.co/jinaai/jina-colbert-v2`

[6] While our work is completely orthogonal to the choice of retriever, we chose this lightweight model that reportedly perform well across several IR benchmarks (Jha et al., 2024).

Evaluation Prompt for Accuracy

# Task:
You are given a Question, a model Prediction, and a list of Ground Truth answers, judge whether the model Prediction matches any answer from the list of Ground Truth answers. Follow the instructions step by step to make a judgement.
1. If the model prediction matches any provided answers from the Ground Truth Answer list, “Accuracy” should be “True”; otherwise, “Accuracy” should be “False.”
2. If the model prediction says that it couldn’t answer the question or it doesn’t have enough information, “Accuracy” should always be “False.”
3. If the Ground Truth is “invalid question,” “Accuracy” is ‘True” only if the model prediction is exactly “invalid question.”

# Output:
Respond with only a single JSON string with an “Accuracy” field which is “True” or “False.”

# Examples:
Question: how many seconds is 3 minutes 15 seconds?
Ground truth: [“195 seconds”]
Prediction: 3 minutes 15 seconds is 195 seconds.
Accuracy: True

Question: Who authored The Taming of the Shrew (published in 2002)?
Ground truth: [“William Shakespeare”, “Roma Gill”]
Prediction: The author to The Taming of the Shrew is Roma Shakespeare.
Accuracy: False

Question: Who played Sheldon in Big Bang Theory?
Ground truth: [“Jim Parsons”, “Iain Armitage”]
Prediction: I am sorry I don’t know.
Accuracy: False

Figure 5: Prompt for evaluating generated answer against ground-truths. Instances classified as ‘False’ are further processed if the model responded with “I don’t know.”

Evaluation Prompt for Sentence-level Precision/Recall

# Task: You are given a Question, a sentence from model Prediction, and the whole Ground Truth answer that may contain several sentences. Judge whether the model Prediction sentence is correctly based on the Ground Truth answer. Follow the instructions step by step to make a judgment.
1. If the content of model prediction is fully implied by the ground truth answer, “Accuracy” should be “True.”
2. If the content of model prediction contains any contradictory or unsupported claim compared to the ground truth answer, “Accuracy” should be “False.”
3. If one of them states “I don’t know the answer,” “Accuracy” should be “True” if and only if the other also states “I don’t know.”

# Output:
Respond with only a single JSON string with an “Accuracy” field which is “True” or “False.”

# Examples:
Question: What is the total amount that Flour Mills of Nigeria (FMN) Plc aims to raise through equity funds over the next three years, and how will these funds be raised?
Ground truth: Flour Mills of Nigeria (FMN) Plc aims to raise up to N40 billion in equity funds over the next three years. These funds will be raised through a rights issue, which will proportionately allot shares to shareholders based on their shareholdings as of a pre-determined date. The board of directors will monitor the capital market conditions to determine the appropriate time to launch the first tranche of the new supplementary issue.
Prediction: The funds will be raised through a rights issue, which will proportionately allot shares to shareholders based on their shareholdings as of a pre-determined date.
Accuracy: True

Question: According to the context, what recognition did Crowne Plaza Resort Salalah receive this year and what natural phenomenon has enhanced the region’s beauty?
Ground truth: Crowne Plaza Resort Salalah was named “Oman’s Leading Resort 2018” by the World Travel Awards this year. The natural beauty of the region has been enhanced by overflowing springs and waterfalls due to the heavy rainfall brought by Cyclone Mekunu, causing the terrains and mountains to turn lush green earlier than expected.
Prediction: The region’s beauty has been enhanced due to the hurricane Mekunu, which blew away all the dirt with strong wind.
Accuracy: False

Question: Who played Sheldon in Big Bang Theory?
Ground truth: I don’t know the answer to that question.
Prediction: I am sorry I don’t know.
Accuracy: True
Question: According to the context, how did Bradley Cooper initially feel about not receiving an Oscar nomination for his directorial debut in “A Star Is Born”?
Ground truth: Bradley Cooper initially felt embarrassed for not receiving an Oscar nomination for his directorial debut in “A Star Is Born,” despite the film garnering critical acclaim and eight nominations, including best picture, actor for Cooper, and actress for Lady Gaga.
Prediction: I don’t know the answer given the passages.
Accuracy: False

Figure 6: Prompt for evaluating sentence-level F_1. To obtain precision, model generated sentence is compared against the ground-truth response. For recall, ground-truth sentence is compared against model-generated response.

C LLM Prompts

We provide prompts used for LLM-as-a-judge evaluation of accuracy (Figure 5) and sentence-level F_1 score (Figure 6).