TO RETRIEVE OR NOT TO RETRIEVE? UNCERTAINTY DETECTION FOR DYNAMIC RETRIEVAL AUGMENTED GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation equips large language models with the capability to retrieve external knowledge, thereby mitigating hallucinations by incorporating information beyond the model's intrinsic abilities. However, most prior works have focused on invoking retrieval deterministically, which makes it unsuitable for tasks such as long-form question answering. Instead, dynamically performing retrieval by invoking it only when the underlying LLM lacks the required knowledge can be more efficient. In this context, we delve deeper into the question, "To Retrieve or Not to Retrieve?" by exploring multiple uncertainty detection methods. We evaluate these methods for the task of long-form question answering, employing dynamic retrieval, and present our comparisons. Our findings suggest that uncertainty detection metrics, such as Degree Matrix Jaccard and Eccentricity, can reduce the number of retrieval calls by almost half, with only a slight reduction in question-answering accuracy.

1 INTRODUCTION

Recently, Large Language Models (LLMs) like ChatGPT OpenAI (2023), Gemini Team et al. (2023), and others are showing impressive strides in tasks across numerous benchmarks Srivastava et al. (2023). This success has been largely owed to their exposure to massive training data and successive fine-tuning of instruction datasets. To increase the helpfulness and decrease the harmfulness of the models, they are being further fine-tuned over preference collections Bai et al. (2022); Ouyang et al. (2022); Rafailov et al. (2024).

Further, Retrieval Augmented Generation (RAG) Lewis et al. (2020); Dhole (2024a); Dhole et al. (2024), in the effort to mitigate hallucinations, enriches these models with domain-specific information and tackles scenarios where the intrinsic knowledge of the base model falls short. By integrating externally retrieved content during the generation phase, RAG enhances the model's ability to produce less hallucinatory and domain-conditioned responses. This approach has been particularly valuable in complex applications such as long-form generation like multi-hop question answering, which often requires multiple retrievals to address a query comprehensively.

However, to optimize the efficiency of RAG, retrieval should only be invoked when necessary — also referred to as conditional retrieval. Previous conditional RAG setups have explored multiple paradigms like low token probabilities Jiang et al. (2023), external classifiers Wang et al. (2023), or low entity popularity Mallen et al. (2023) as indicators of the LLMs' knowledge gaps. However, most of these methods fall short in either approximating knowledge gaps of the LLMs or lacking the ability to invoke retrieval dynamically.

On the other hand, with the potential of LLMs to hallucinate, there has been an increasing interest in **uncertainty detection** methods to gauge LLMs' confidence in their outputs Fadeeva et al. (2023). Unlike traditional methods that rely on rigid heuristics or external classifiers, uncertainty detection leverages the inherent variability in LLM-generated responses to estimate confidence dynamically.

For instance, semantic sets-based UD approaches Lin et al. (2023) group responses based on meaning, and use the number of clusters to directly reflect the level of uncertainty — with greater variability signaling higher uncertainty. Similarly, spectral methods using eigenvalue Laplacians quantify

response diversity by identifying strong or weak clustering patterns in pairwise similarity graphs. These approaches align with the probabilistic nature of LLMs as well as adaptively gauge uncertainty based on output coherence, making them more robust against adversarial or ambiguous inputs.

In this work, we evaluate if such uncertainty detection methods can indeed enhance the reliability of conditionally invoking retrieval, by measuring its impact on a downstream task of multi-hop question answering.

In that regard, we resort to a conditional RAG system and employ numerous uncertainty detection metrics to test the need for invoking retrieval. Our RAG system performs forward-looking active retrieval in the style of Jiang et al. (2023) Jiang et al. (2023).

Specifically, we contribute the following:

- We design a retrieval augmented generation with dynamic retrieval
- We perform an exhaustive analysis of various conditions from the "uncertainty quantification" literature to gauge the best strategy to dynamically retrieve during generation
- Based on the results, we present insights for future research

Our insights are useful to gauge whether uncertainty detection methods can help improve the efficiency of RAG.

2 Related Work

Here, we summarise some of the related work on uncertainty quantification and some active RAG efforts.

There has been a lot of recent work on uncertainty quantification of white box and black box NLG models. Lin et al. (2023) showed that along with their generations, GPT-3 can output a verbalized form of the uncertainty, viz. "high confidence" or "85% confidence". Kadavath et al. (2022) show that models can be made to sample answers and then made to self-evaluate the probability of P(True). Kuhn et al. (2023) recently proposed to compute the semantic entropy by considering the equivalence relationships amongst generated responses.

We now describe the tasks and datasets used in our analysis along with the UD approaches employed.

3 Tasks and Datasets

We conduct experiments on the 2WikiMultihopQA dataset Ho et al. (2020), a multi-hop open domain question answering (QA) dataset that tests the reasoning and inference skills of question-answering models. Questions in this dataset generally require two steps of reasoning to deduce the final answer, and the information for each step of reasoning can be obtained through referencing external information viz., Wikipedia passages.

4 Approach

We now describe our uncertainty-aware, retrieval-augmented generation in the following two subsections.

4.1 Uncertainty Evaluation of Future Sentence

Given a query $\mathbf{q}$, a retriever $\mathbf{R}$, a text generator $\mathbf{G}$, and a black box uncertainty estimation function $\mathbf{U}$, and partially generated sequence $t_{<i}$ until time step i, – we first generate a temporary sentence t_n in the style of FLARE Jiang et al. (2023).

We use a prompt template $\mathbf{P}$, which could take the form of a zero-shot or a few-shot instruction. This instruction takes as input the query, zero or more retrieved documents $d_1 \dots d_k$, and the answer tokens generated until now. Here, we use t_i to represent the i^{th} temporary sentence and $y_{<i}$ to represent

all the initialised and generated sentences $\{0 \dots (i-1)\}$. t_i is first obtained without performing retrieval:

$$t_i = \mathbf{P}\{\mathbf{q}, \dots, y_{i-1}\} \tag{1}$$

During generation, we evaluate the uncertainty of this temporary sentence t_n to gauge if the generator needs more information. If the uncertainty $\mathbf{U}(t_n)$ exceeds a threshold $\theta_{\mathbf{U}}$, the model is not certain and may lack the necessary knowledge to provide an accurate answer. The next sentence y_i is then computed by appending retrieved information to the model context:

$$y_i = \begin{cases} \mathbf{P}\{d_1, \dots, d_k, \mathbf{q}, \dots, y_{i-1}\} & \text{if } \mathbf{U}(t_i) > \theta_{\mathbf{U}} \\ \mathbf{P}\{\mathbf{q}, \dots, y_{i-1}\} & otherwise \end{cases} \tag{2}$$

where $d_1 \dots d_k$ are obtained from a retrieval system ϕ.

$$d_1 \dots d_k := \phi(\mathbf{q}) \tag{3}$$

4.2 Sequence Level Uncertainty Evaluation Measures

We resort to 5 recently introduced sequence-level uncertainty evaluation measures. Each of them work in a black box manner without requiring information regarding the model parameters.

The high-level strategy of all the methods is the same. Given an input x, first generate n responses through some generator G and then compute pairwise similarity scores of each of the n responses with each other. Using these similarity values, compute an uncertainty estimate $U(x)$ or a confidence score.

- **Semantic Sets**: In the black-box approach of kuhn2023semantic, the authors propose to compute semantic sets i.e. groups of responses that are close together in meaning. These semantic sets of equivalence subsets are computed using a Natural Language Inference (NLI) classifier. Here, the number of semantic sets can be regarded as an uncertainty estimate as when the responses differ in meaning, the number of groups increases.
- **Eigen Value Laplacian**: defines the uncertainty estimate by capturing the essence of spectral clustering. First, an adjacency matrix is created from the pairwise similarities of responses. Then the matrix is partitioned into clusters, where each cluster corresponds to a distinct "meaning" or category within the responses. The eigenvalues close to one indicate strong cluster formations, thus contributing less to the uncertainty estimate, while those further from one suggest weaker clustering or more diffuse distributions of responses, hence increasing the uncertainty estimate.
  The degree matrix of the adjacency graph is also used to compute the uncertainty estimate **?**. A node that is well-connected to other nodes, might be less uncertain. We use two similarity metrics for computing the degree matrix.
- **Degree Matrix (Jaccard Index)**: The Jaccard similarity is a light-weight metric where sentences or passages are treated as sets of words, and similarity between responses is computed by taking the fraction of the intersection of the two sets and the union of the two sets.
- **Degree Matrix (NLI)**: Here, the similarity between responses is computed through classifying entailment relations amongst them. A classifier predicts whether a pair of responses contradict, entail, or are neutral to each other.

Uncertainty Estimator	Trigger Retrieval When	Retrieval Query	#examples	#search	#steps	f1
Always Retrieve	U $\geq$ 0	Temporary Sentence	25	4.60	3.60	0.552
Always Retrieve		Sub-Query	25	5.00	4.00	0.538
FLARE-Instruct	"...[Search"		25	4.80	3.80	0.531
Degree Matrix Jaccard	U $>$ 0.4	Sub-Query	24	1.46	3.67	0.593
Eccentricity	U $>$ 2	Sub-Query	22	2.23	4.05	**0.605**
Semantic Sets	U $>$ 2	Sub-Query	23	2.52	4.09	0.411
Degree Matrix NLI	U $>$ 0.5	Sub-Query	24	2.25	4.00	0.535

Table 1: Performance Metrics over a smaller seed set

Uncertainty Estimator	Trigger Retrieval When	#search	#steps	ret ratio	correct	incorrect	f1
Always Retrieve	Always	4.63	3.63	1.32	0.493	0.493	0.578
		4.61	3.61	1.33	0.52	0.467	0.594
		4.61	3.61	1.33	0.493	0.493	0.571
							0.581
Degree Matrix Jaccard	U $>$ 0.4	1.80	3.61	0.57	0.453	0.533	0.538
		1.92	3.60	0.61	0.44	0.547	0.525
		1.85	3.61	0.57	0.419	0.568	0.508
							0.524
Eccentricity	U $>$ 2	2.17	3.60	0.64	0.44	0.547	0.525
		2.25	3.63	0.67	0.467	0.533	0.565
		2.23	3.63	0.64	0.507	0.493	0.594
							0.561

Table 2: Performance Metrics for Different Uncertainty Estimators for 75 examples.

4.3 Subquery Generation for Retrieval

We resort to retrieving relevant knowledge to account for the information that the model is lacking to answer the question. FLARE Jiang et al. (2023) generates a retrieval query for the missing entity in the temporary sentence by using the sentence with the low probability token removed or by prompting an external question generator to generate a question for the missing entity as the answer. We generalize this by instead prompting the model to generate a subquery to figure out the missing information needed to answer the user query in an open-ended manner.

We define a subquery generator $\mathbf{S_Q}$ which takes in as input few-shot exemplars of subqueries, the current user query $\mathbf{q}$, and the current partial answer sentences uttered in chain-of-thought Wei et al. (2022) fashion. It seeks to generate subqueries to get a specific piece of information not generated in the partial answer sentences but is needed to answer $\mathbf{q}$. Once this subquery is generated, we use this subquery to retrieve additional passages from the external retriever $\mathbf{R}$. These passages are then appended to the user input, and the generation continues.

For instance, for the question, "Which film has the director who died first, Promised Heaven or Fire Over England?", and the partially generated answer, "The film Promised Heaven was directed by Eldar Ryazanov. Fire Over England was directed by William K. Howard. Eldar Ryazanov died on November 30, 2015.", we expect the model to generate a subquery, "When did William K. Howard die?".

5 Setup

The generator used in all experiments was GPT-3 (davinci-002) Brown et al. (2020), and the retriever employed was BM25 through PyTerrier Macdonald et al. (2021); Dhole (2024b). The base code used for conducting the experiments and computing the metrics presented in the tables was obtained from the active RAG setup by Jiang et al. (2023). For uncertainty detection, we resort to the Fadeeva et al. (2023)'s LM-Polygraph library.

Since running GPT-3 (davinci-002) along with many of the uncertainty detection metrics could be expensive to run (due to making multiple calls), we first perform a run for a small seed set of 25 queries across all metrics and then choose the 3 best metrics for a rerun across a larger set of 75 examples. We perform each run three times.

6 Results

We now present the results in Tables 1 and 2 for the smaller and the larger sets respectively.

The baseline method where retrieval was always invoked yielded an F1 score of **0.552** when using temporary sentences as retrieval queries and **0.538** when subqueries were generated for retrieval but required most number of retrieval operations.

Triggering retrieval, when uncertainty computed through **Eccentricity** i.e. **U ¿ 2**, led to the highest F1 score of **0.605**, with a lesser number of search operations. This approach balanced retrieval efficiency and task performance better than other methods. It required half the number of search operations than an Always Retrieve approach. **Semantic Sets'** innovative clustering approach performed poorly, with an F1 score of **0.411**. Using entailment-based similarity to compute uncertainty via the **Degree Matrix NLI** measure achieved an F1 score of **0.535**, comparable to the baseline. The lightweight **Degree Matrix (Jaccard)** necessitated the least number of retrieval operations to perform better than an Always Retrieve baseline.

Table 2 presents additional performance metrics over a larger set of 75 examples. Notably, the **Eccentricity** method consistently demonstrated the best balance between retrieval efficiency and performance, achieving an average F1 score of **0.561** across different experimental runs, while reducing unnecessary retrievals compared to the baseline.

Degree Matrix (Jaccard) performed slightly worse in F1 score (**0.524**) but depended on retrieval the least indicating its potential for applications where minimizing retrieval costs is crucial.

In contrast, the **Always Retrieve** approach performed better than both conditional retrieval approaches but necessitated almost twice the number of retrieval calls.

7 Conclusion

Our experiments demonstrate that dynamic retrieval, guided by uncertainty detection, improves the efficiency of retrieval-augmented generation systems, making it useful where retrieval can be expensive to compute. Among the methods tested, **Eccentricity-based uncertainty detection** emerged as the best-performing approach, offering the highest F1 score with a moderate number of retrieval steps and searches. This method effectively balances retrieval efficiency with task performance.

The **Degree Matrix (Jaccard)** method also showed promising results, particularly in reducing retrieval costs while maintaining reasonable performance. Conversely, methods such as **Semantic Sets** and **FLARE-Instruct** underperformed, highlighting the need for more reliable uncertainty estimators.

Although some black-box uncertainty detection methods require multiple runs of generation, which can be costly, always retrieving may be preferable in RAG applications where lightweight retrieval methods like BM25 suffice. This is also evident from the results on the larger set.

Besides, we feel that uncertainty detection might become more mainstream as the propensity for hallucination in LLMs increases, and as end applications demand more confidence and interpretability Dhole et al. (2024) in their outputs making uncertainty detection a necessity. Our work focuses on exploiting uncertainty detection for RAG, especially where retrieval can be expensive like the usage of heavy and composite retrieval systems employing numerous components like reformulation, dense retrieval Santhanam et al. (2021), reranking, etc.

8 Ethical Considerations

When evaluating large language models (LLMs), it is essential to adopt a sociotechnical perspective Dhole (2023), acknowledging that their outputs are influenced by both social contexts and technical design choices. Proper safeguards should be in place to mitigate biases and prevent the generation of harmful or toxic content. Furthermore, the uncertainty detection approaches we employed rely on estimations derived from various neural network computations, which are inherently shaped by the data on which the models are trained. Consequently, it is critical to thoroughly test uncertainty detection methods to ensure they meet the requirements of the intended applications.

Despite these precautions, there remains a possibility that some approaches may misrepresent the level of certainty, as no method is flawless. Therefore, ongoing evaluation and refinement of uncertainty detection mechanisms are necessary to minimize inaccuracies and potential misinterpretations.

REFERENCES

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

Kaustubh Dhole. Large language models as sociotechnical systems. In *Proceedings of the Big Picture Workshop*, pp. 66–79, 2023.

Kaustubh Dhole. Kaucus-knowledgeable user simulators for training large language models. In *Proceedings of the 1st Workshop on Simulating Conversational Intelligence in Chat (SCI-CHAT 2024)*, pp. 53–65, 2024a.

Kaustubh D Dhole. Pyterrier-genrank: The pyterrier plugin for reranking with large language models. *arXiv preprint arXiv:2412.05339*, 2024b.

Kaustubh D. Dhole, Kai Shu, and Eugene Agichtein. Conqret: Benchmarking fine-grained evaluation of retrieval augmented argumentation with llm judges, 2024.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-polygraph: Uncertainty estimation for language models. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.41. URL https://aclanthology.org/2023.emnlp-demo.41.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://www.aclweb.org/anthology/2020.coling-main.580.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7969–7992, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.495. URL https://aclanthology.org/2023.emnlp-main.495.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

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

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2023.

Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In *Proceedings of the 30th acm international conference on information & knowledge management*, pp. 4526–4533, 2021.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546.

OpenAI. Gpt-4 technical report, 2023.

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

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

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488*, 2021.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10303–10315, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.691. URL https://aclanthology.org/2023.findings-emnlp.691.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.