# EVALUATING THE ROBUSTNESS OF RETRIEVAL PIPELINES WITH QUERY VARIATION GENERATORS

*by*
*KRISTA BRADSHAW*

School of Information Technology and Electrical Engineering,
The University of Queensland.

Project Proposal Draft

MARCH & 2023.

# Contents

# Chapter 1

# Introduction

## 1.1 Project Definition

This thesis is a reproduction of the study by Penha, Câmara and Hauff titled "Evaluating the robustness of retrieval pipelines with query variation generators" [1]. It aims to demonstrate an in-depth understanding of Large Language Models (LLMs) and their applications in powering search engines, specifically how query variations impact robustness. The topic can be summarised by the following research question: *What effect do query variations have on the robustness of retrieval pipelines?* This question is important because it addresses a fundamental section of research on retrieval pipelines and their ability to handle variations in user-generated queries. There is ongoing interest in improving the performance and robustness of these pipelines with many researchers actively working to address the challenges associated with variations in user query semantics. This study can help to identify potential limitations and areas for improvement in the design and implementation of these pipelines. The results of this study can also provide insights into how to better optimise retrieval systems to meet the needs and expectations of users.

## 1.2 Aims

The high level aims of this project are centred around the goals and objectives of Penha et al.'s study [1]. They are as follows:

- To gain an in-depth understanding of the previous literature in this field and identify gaps to be further investigated;
- To reproduce the experiments conducted in [1], document the process, and investigate the findings in comparison to those of the original study;
- To expand on the original experiments by including the DL-typo dataset [2] and other relevant datasets; and
- To identify and provide insights into the strengths and weaknesses of information retrieval tasks, and to highlight the need for continued research in this area.

# Chapter 2

# Background

## 2.1 Information Retrieval

Information Retrieval (IR) is the process of retrieving relevant information from large collections of data, typically in the form of text documents or web pages. IR systems use various techniques and algorithms to analyse and index large datasets, such as search engine indexes, document repositories or databases . When a user enters a search query, the IR system searches through the indexed data and returns a list of documents or web pages in order of relevance.

### 2.1.1 Large Language Models

Large Language Models (LLMs) are machine learning models utilised in Natural Language Processing (NLP) applications that are trained on copious collections of text to generate natural language text or predictions [1]. LLMs have diverse real-world applications, including chatbots, language translation services, and voice assistants. Furthermore, they have facilitated new applications in content generation, language modelling, and text summarisation [3]. LLMs are a valuable tool in NLP by advancing machines' ability to comprehend, process, and produce natural language text.

Retrieval pipelines are a series of processing steps used in IR systems to retrieve relevant information from a large collection of data, such as a document database or a search engine index. Retrieval pipelines typically involve the following steps:

1. Query processing: This involves processing the query to identify relevant terms, expanding the query with synonyms, and applying other query optimisation techniques.

2. Document indexing: This step involves indexing the documents in the collection, typically using techniques such as token embedding, to identify the most relevant documents.

3. Ranking: The ranked documents are then sorted by relevance, typically using models such as BM25 and monoBERT.

4. Presentation: The final step involves presenting the search results to the user in a user-friendly format, such as a list of documents with summaries [4, 5].

Retrieval pipelines can also include additional steps, such as filtering or classifying the search results based on domain-specific criteria or using machine learning techniques to improve relevance comparisons [6].

### 2.1.2 Datasets

In IR, large collections of text data are commonly used for indexing and searching. These datasets are publicly available and serve as benchmarks for IR research. Some notable datasets used are:

- TREC (Text Retrieval Conference) collection is a collection of text documents that have been used as a benchmark in IR research for several decades. It contains a wide range of document types, including news articles, scientific papers, and web pages [7].
- UQV100 is a benchmark dataset used for evaluating the effectiveness of information retrieval models with a focus on query variations. It contains a set of 100 topics, which are information needs or search queries, and corresponding sets of relevant documents [8].
- MS MARCO (Microsoft MAchine Reading COmprehension) is a large-scale dataset that contains over 1 million queries sampled from Bing's search logs, with human-generated relevance judgments [9].

### 2.1.3 Common Models

Some common models are as follows:

- BM25 (Best Match 25) is a traditional retrieval model used for ranking documents based on their query relevance. It uses a similarity function that considers term frequency and document length [10].
- BERT (Bidirectional Encoder Representations from Transformers) is a transformer based LLM which can be fine-tuned for various NLP tasks. It is pre-trained using next sentence prediction and masked language modelling on large datasets [11]. There exists many different versions and expansions of BERT which cater to specific situations such as large collections and multi-language.
- EPIC (Efficient Passage-based Interactive Cross-lingual Information Retrieval) is a transformer based retrieval model used for cross-lingual retrieval. It uses an efficient passage-based approach to retrieve relevant documents for a given query in a different language [12].

- KNRM (Kernel-based Neural Ranking Model) is an end-to-end neural ranker designed to target the limitations of traditional models by taking into account the semantic similarity between documents and queries using a kernel function [13].

## 2.2 Natural Language Processing

### 2.2.1 Ranking

Ranking refers to the process of ordering a set of text documents or search results based on their relevance to a given query or topic. Ranking algorithms typically use various features to assign scores to each document. Then those documents are ranked in descending order of their score, having the most relevant documents appearing first [14]. The goal of ranking is to improve the performance of IR systems by ensuring that the most relevant documents appear to the user first. Ranking is a critical component of many applications, such as search engines, chatbots, and question answering systems.

### 2.2.2 Training

Training refers to the process of teaching a model to perform specific NLP tasks, such as text classification, sentiment analysis, or translation [11]. The training process involves feeding large amounts of text data into the model, along with corresponding target labels. The model then learns to recognise relationships and patterns based on the labelled examples provided [15]. The training process typically involves iterating over the data multiple times, using optimisation techniques to adjust the model's parameters and improve its performance [15]. The ultimate goal of training is to develop models that can accurately and efficiently process natural language text for a vast range of applications.

## 2.3 Effectiveness Evaluation

The effectiveness of LLMs is evaluated by measuring performance on specific NLP tasks. Typically, this evaluation is conducted offline using a benchmark dataset and a well-defined task, such as language translation, sentiment analysis or text classification.LLMs are commonly evaluated using a variety of datasets and tasks, including the BM25 benchmark and other similar datasets. The results of evaluations like these can be used to compare the effectiveness of different LLMs and to identify areas that need strengthening [1].

Online evaluation is a technique to assess the effectiveness of an IR system in real-time using real-world applications. Real-world applications involve deploying LLM systems in actual settings and measuring their performance and impact on user outcomes, such as improved efficiency. In contrast to offline evaluation, which is performed on pre-existing data sets, online evaluation measures a system's performance by observing how users interact with it during live usage.

In addition to data focused evaluations, the effectiveness of LLMs can also be assessed through user studies. User studies typically involve asking users to perform specific tasks or interact with a system and collecting data on their satisfaction and experience. There are several different types of user studies, including surveys, and usability testing.

### 2.3.1 Measures

The effectiveness is evaluated in terms of certain measures such as accuracy, recall, and precision, as described below. These measures allow researchers to weigh a models' ability to correctly identify and classify different types of language data [16].

- Accuracy: This metric measures the percentage of correctly classified instances out of all instances in the dataset.
- Precision: This metric measures the proportion of true positives (correctly classified documents) verses all documents classified as positive.
- Recall: This metric measures the proportion of true positives out of all documents that are actually positive.
- fMeasure: This metric is the harmonic mean of recall and precision, it is commonly used when the dataset is imbalanced and to emphasise the importance of recall over precision.
- MAP (Mean Average Precision): Unlike the above measures, accounts for all rankings of relevant documents by calculating the mean of the average precision scores across all queries.

## 2.4 Query Variation

Query variation refers to the different expressions or ways in which users can convey a specific information need or search query [5]. For example, a user might use different word order or sentence structure to express the same need. Consider the following two queries related to a user's interest in baking a cake: "How to bake a cake from scratch?" "Homemade cake recipe?" Both queries express the same information need while having slight variation. Other examples of query variations may demonstrate specialisation, aspect change, misspelling, natural language, or paraphrasing [1].

To handle query variation in IR systems, various techniques such as query suggestion, reformulation, or expansion can be employed. Query expansion is the process of adding synonyms or extra terms to the original query to expand its scope and retrieve more relevant documents. Query reformulation is when the original query is augmented into a new query that better reflects the information need [17]. Addressing query variation is crucial in information retrieval since it can significantly impact the effectiveness of a search system. By understanding and accounting for the diverse ways in which users express their information needs, search systems can retrieve more relevant documents and offer an improved user experience.

# Chapter 3

# Literature Review

Query variations can arise due to various factors such as ambiguity in user search terms, different user intentions, or variations in the context of the search. By evaluating the existing literature on query variation generators and their place within LLMs, this review aims to identify the current state of research in this area and highlight any gaps or areas for future research.

## 3.1 Large Language Models

In recent studies there has been a focus on improving the performance and efficiency of LLMs in NLP tasks, particularly in the domain of text generation and understanding. One key trend in this field is the use of pre-trained models, such as GPT [3] and BERT [11], which achieve state-of-the-art results on a various NLP tasks. Studies have also investigated the use of transfer learning techniques, such as fine-tuning [11] and domain adaptation [14], to improve the performance of LLMs in specific contexts. Future research is expected to explore language models and their use in more complex and diverse NLP tasks, as well as the development of new methods to improve the efficiency and scalability of these models [17, 18, 19].

### 3.1.1 Improvement of Retrieval Pipelines

Recent literature on retrieval pipelines has centered around improving the effectiveness and efficiency when handling large-scale datasets and diverse ranges of queries [1, 14, 20]. The main themes in this field of research include the use of deep learning techniques, including deep reinforcement learning and neural networks [1, 17, 21], the investigation of query expansion and query prediction techniques [5, 6, 20], and the development of novel evaluation metrics, such as diversity measures, to evaluate the relevance and novelty of search results [16, 20].

Overall, recent literature on retrieval pipelines aims to develop more effective and efficient methods for information retrieval that can enhance search quality and user satisfaction. Zendel et al.'s [5] work is a novel investigation into the relationship between query and information need, with their approach to query performance prediction (QPP) tasks outperforming the baseline. While there is an abundance of research into improving these pipelines there is still ways to go, there are studies which purposefully target the vulnerabilities to highlight weaknesses [21, 22]. Gao et al.'s [21] work in this field presents a novel framework designed to generate queries formulated to trick models into classifying them incorrectly. Their results show a drastic decrease in effectiveness.

### 3.1.2 Evaluation

Recent literature on the evaluation of LLMs has emphasised the need for more diverse and robust evaluation methods [1, 14, 17, 20]. Traditionally, LLMs have been evaluated using metrics such as accuracy, which measures the model's ability to correctly predict the classification of text. However, it has been theorised that in real-word applications, this metric may not appropriately reflect the model's performance [16, 23]. Moffat et al. [20] introduced the C/W/L framework in an attempt to better model true user behaviour by considering the probability that a user, while looking at a document, will choose to continue onto the next. This study showed that by considering an "expected goal of search", more accurate predictions of a user's need. Additionally, Ribeiro et al. [23] proposed a novel evaluation model that addresses the limitations of conventional evaluation measures and methods. Their experiments revealed previously undetected shortcomings in NLP models that had undergone extensive testing prior.

## 3.2 Query Variation

### 3.2.1 Benchmark Datasets

Many recent studies have proved the importance of the presence of query variations in benchmark datasets [1, 5, 6]. This is important for several reasons: it makes the datasets more realistic and representative of the types of queries users use in real-world situations, including query variations ensures that benchmark datasets cover a wide range of queries and information needs, and it allows for better evaluation of the effectiveness of IR systems in handling variations in user queries [1, 2, 5]. Some datasets have been generated specifically to cater to these needs such as UQV100 [8] , TREC [7] and recently, DL-typo [2]. The study by Zhuang et al. specifically aimed to tackle the notion that current dense retrievers struggle to perform with unusual queries. Furthermore, as discovered by Penha et al. [1] following their investigation into retrieval pipelines robustness when using datasets with query variations, using these benchmark datasets can stimulate improvements in IR systems by highlighting areas of weaknesses. These studies have initiated the advancement of this field, and this experiments of this study intends to carry on their research. Overall, the inclusion of query variations in benchmark datasets is critical for evaluating and enhancing the performance of IR systems in handling variations in user queries and information needs.

### 3.2.2 Effect on Performance

Research on query variation is crucial because it is well-established that such variations have a detrimental effect on the efficacy of LLMs to date. Many studies have shown that text variations are the cause of under-performing IR systems [2, 4, 23]. The study by Penha et al. [1] modeled 7 approaches to ranking with 2 datasets and their findings showed that when faced with 4 different types of query variations, the effectiveness dropped an average of 20%. This proves the importance of further study in this field and ways to improve on the existing models. Further, Lu et al.'s [6] study on relevance modelling with multiple queries shows that the utilisation of query variations substantially outperforms that of a single query. Their experiments involved performing fusion at the term, query, and document level which is a new concept to this field and shows promising results. By incorporating query variations, LLMs are better able to capture the nuances and complexities of user queries and information needs, resulting in more accurate and relevant search results, as provide in several recent studies [1, 2, 5, 6]. Zuccon et al.'s [18] study focused on examining the role of query variations in comparing system effectiveness, and proposed a framework that explicitly incorporates query variations. Unlike similar studies, their analysis considers not only the mean effectiveness of the system, but also its variance across different query variations and topics. The findings reveal a significant impact of query variations on the comparison of different systems.

### 3.2.3 Methods of Generation

Recent research has explored the use of method to generate queries that better match user intent and handle variations such as data augmentation, fusion, query reformulation, and adversarial training [6, 17, 19]. Namely, Bailey et al. [17] proves that query formulation is critical to effectiveness in their study of query variability in LLM evaluation. Benham et al. [19] also investigates building query variations using fusion and weighted random sampling process and it resulted in their retrieval effectiveness being competitive with the state-of-the-art.

Overall, recent literature on query variation in IR highlights the critical role of datasets, the potential impact on performance, the promising approach of query generation, and the need for more appropriate evaluation metrics. Accounting for query variation is essential for developing effective systems that can understand and respond to users' information needs.

# Chapter 4

# Project Plan

## 4.1 Project Plan

To ensure this project progresses smoothly and stays on track, it can be divided into several milestones which outline important steps to be undertaken to achieve ordered goals.

### 4.1.1 Milestones

**Milestone 1 – Project Initiation & Initial Testing**

This milestone focuses on gathering and preparing all necessary resources and beginning foundational testing. Steps to complete Milestone 1 include:

1. Create GitHub repository for the project codebase and add supervisors
2. Gain access to, install and prepare all required models
3. Gain access to, download and prepare all required datasets
4. Begin initial testing to ensure models are working as expected in local environment
5. Document the results and note any unexpected outcomes

This milestone is expected to commence immediately after the submission of Deadline 1 (See *4.1.2* for breakdown of Deadlines) and conclude after roughly 1 month.

**Milestone 2 – Finalised Project Plan**

This milestone focuses on nailing down the project plan in preparation for submission of the final Project Proposal and preparing to conduct experiments. Steps to complete Milestone 2 include:

1. Based on initial testing, decide on how to best expand on the original experiments
2. Finalise written project proposal
3. Gain access to, download and prepare any new models and datasets

4. Backup any and all collected data thus far on cloud service in case of hard drive failure

This milestone is expected to commence early April and conclude no later than the submission of Deadline 2.

### Milestone 3 – Replication of Experiments

This milestone focuses on conducting all experiments as per [1] and beginning to analyse the results in preparation for expanding the experiments further. Steps to complete Milestone 3 include:

1. Replicate the experiments conducted in [1]
2. Document all processes, outcomes and results
3. Analyse collected results against the original results and note discrepancies
4. Based on these results, potentially revisit how to best expand on the original experiments

This milestone is expected to commence around the submission of Deadline 2 and conclude within 3 months.

### Milestone 4 – Expansion of Experiments

This milestone focuses making adjustments to the resources and expanding on the original experiments to address gaps in this field of study. Steps to complete Milestone 4 include:

1. Create copies of datasets and make necessary changes
2. Conduct experiments on new datasets
3. Conduct any new experiments
4. Document all processes, outcomes and results
5. Analyse collected results

This milestone is expected to commence around the submission of Deadline 3 and conclude within 3 months.

### Milestone 5 – Finalised Thesis Project

This milestone focuses on finalising all experiments and analysing of data in preparation for submission of the written thesis report. Steps to complete Milestone 5 include:

1. Compile all relevant experiment results into a palatable format
2. Critically analyse all results and compare to that of [1]

3. Conduct any final experiments, if necessary
4. Document all progress, results and implications for further work in final written thesis

This milestone is expected to durate the final few months of the project.

### 4.1.2 Timeline

The milestones and deadlines involved in this project, in chronological order, are outlined as follows and then visualised in Figure 4.1 below:

1. Deadline 1 – Proposal Draft – 23rd March
2. Milestone 1 – Project Initiation & Initial Testing – Mid April
3. Milestone 2 – Finalised Project Plan – Late April
4. Deadline 2 – Proposal – 27th April
5. Milestone 3 – Replication of Experiments – Late July
6. Deadline 3 – Progress Seminar – 12th May
7. Milestone 4 – Expansion of Experiments – Late September
8. Milestone 5 – Finalised Thesis Project – 6th November
9. Deadline 4 – Demonstration – 20th October
10. Deadline 5 – Thesis Submission – 6th November

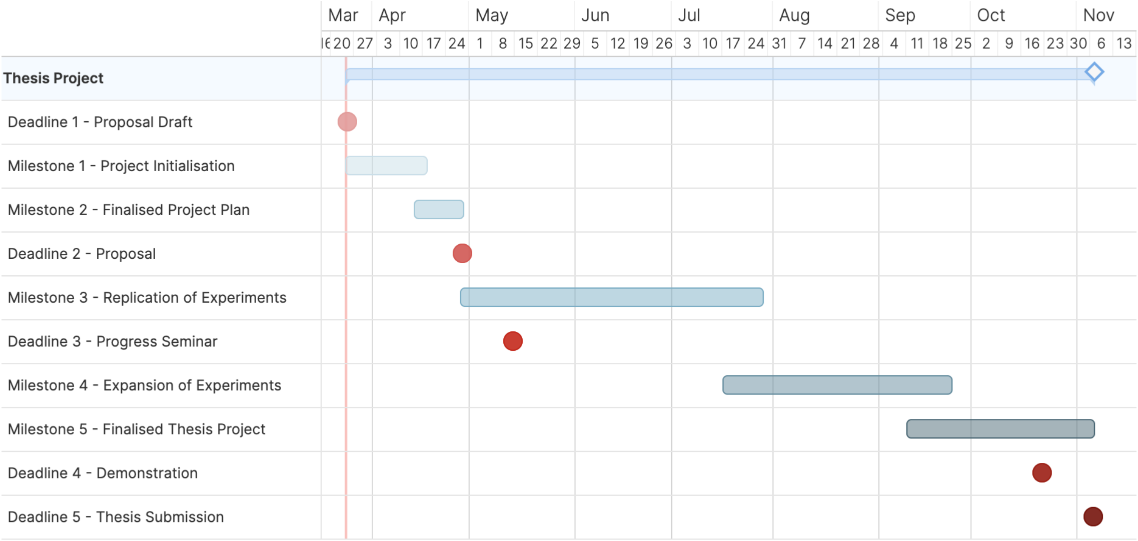

Figure 4.1: Gantt Chart

# Bibliography

[1] G. Penha, A. Câmara, and C. Hauff, “Evaluating the robustness of retrieval pipelines with query variation generators,” in *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*, pp. 397–412, Springer, 2022.

[2] S. Zhuang and G. Zuccon, “Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos,” in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1444–1454, 2022.

[3] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.

[4] X. Chen, B. He, K. Hui, L. Sun, and Y. Sun, “Dealing with textual noise for robust and effective bert re-ranking,” *Information Processing & Management*, vol. 60, no. 1, p. 103135, 2023.

[5] O. Zendel, A. Shtok, F. Raiber, O. Kurland, and J. S. Culpepper, “Information needs, queries, and query performance prediction,” in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 395–404, 2019.

[6] X. Lu, O. Kurland, J. S. Culpepper, N. Craswell, and O. Rom, “Relevance modeling with multiple query variations,” in *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 27–34, 2019.

[7] C. Buckley and J. A. Walz, “The trec-8 query track.,” in *TREC*, 1999.

[8] P. Bailey, A. Moffat, F. Scholer, and P. Thomas, “Uqv100: A test collection with query variability,” in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 725–728, 2016.

[9] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng, “Ms marco: A human generated machine reading comprehension dataset,” *choice*, vol. 2640, p. 660, 2016.

[10] S. E. Robertson and S. Walker, “Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval,” in *SIGIR’94: Proceedings of the*

*Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pp. 232–241, Springer, 1994.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] S. MacAvaney, F. M. Nardini, R. Perego, N. Tonellotto, N. Goharian, and O. Frieder, "Expansion via prediction of importance with contextualization," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 1573–1576, 2020.

[13] C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power, "End-to-end neural ad-hoc ranking with kernel pooling," in *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pp. 55–64, 2017.

[14] C. Wu, R. Zhang, J. Guo, Y. Fan, and X. Cheng, "Are neural ranking models robust?," *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 1–36, 2022.

[15] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[16] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence: 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006. Proceedings 19*, pp. 1015–1021, Springer, 2006.

[17] P. Bailey, A. Moffat, F. Scholer, and P. Thomas, "User variability and ir system evaluation," in *Proceedings of The 38th International ACM SIGIR conference on research and development in Information Retrieval*, pp. 625–634, 2015.

[18] G. Zuccon, J. Palotti, and A. Hanbury, "Query variations and their effect on comparing information retrieval systems," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 691–700, 2016.

[19] R. Benham, J. S. Culpepper, L. Gallagher, X. Lu, and J. M. Mackenzie, "Towards efficient and effective query variant generation.," in *DESIRES*, pp. 62–67, 2018.

[20] A. Moffat, P. Bailey, F. Scholer, and P. Thomas, "Incorporating user expectations and behavior into the measurement of search effectiveness," *ACM Transactions on Information Systems (TOIS)*, vol. 35, no. 3, pp. 1–38, 2017.

[21] J. Gao, J. Lanchantin, M. L. Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*, pp. 50–56, IEEE, 2018.

[22] S. Garg and G. Ramakrishnan, "Bae: Bert-based adversarial examples for text classification," *arXiv preprint arXiv:2004.01970*, 2020.

[23] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of nlp models with checklist," *arXiv preprint arXiv:2005.04118*, 2020.