On the Robustness of Generative Information Retrieval Models: An Out-of-Distribution Perspective

Yu-An Liu[1,2][0000−0002−9125−5097], Ruqing Zhang[1,2][0000−0003−4294−2541], Jiafeng Guo[1,2⋆][0000−0002−9509−8674], Changjiang Zhou[1,2][0009−0000−0005−9465], Maarten de Rijke[3][0000−0002−1086−0202], and Xueqi Cheng[1,2][0000−0002−5201−8195]

[1] CAS Key Lab of Network Data Science and Technology, ICT, CAS
[2] University of Chinese Academy of Sciences
[3] University of Amsterdam
{liuyuan21b, zhangruqing, guojiafeng, zhouchangjiang23s, cxq}@ict.ac.cn
m.derijke@uva.nl

Abstract. Generative information retrieval methods retrieve documents by directly generating their identifiers. Much effort has been devoted to developing effective generative information retrieval (IR) models. Less attention has been paid to the robustness of these models. It is critical to assess the out-of-distribution (OOD) generalization of generative IR models, i.e., how would such models generalize to new distributions? To answer this question, we focus on OOD scenarios from four perspectives in retrieval problems: (i) *query variations*; (ii) *unseen query types*; (iii) *unseen tasks*; and (iv) *corpus expansion*. Based on this taxonomy, we conduct empirical studies to analyze the OOD robustness of representative generative IR models against dense retrieval models. Our empirical results indicate that the OOD robustness of generative IR models is in need of improvement. By inspecting the OOD robustness of generative IR models we aim to contribute to the development of more reliable IR models. The code is available at https://github.com/Davion-Liu/GR_OOD.

Keywords: Generative retrieval · Robustness · Out-of-distribution.

1 Introduction

With the development of representation learning techniques [8], considerable progress has been made in dense retrieval based on the "index-retrieve" pipeline [5, 13, 34]. Information retrieval (IR) approaches based on the index-retrieve pipeline may suffer from a large memory footprint and difficulties in end-to-end optimization. Recently, a generative information retrieval paradigm has been proposed [35]. In this paradigm, different components for indexing and retrieval are fully parameterized with a single consolidated model. Specifically, a sequence-to-sequence (seq2seq) learning framework is employed to directly predict the identifiers of relevant documents (docids) with respect to a given query.

⋆ Corresponding author

Current research on generative IR is often conducted in a homogeneous and narrow setting. That is, work on generative IR often assumes that the training and test examples are independent and identically distributed (IID). Under the IID assumption, the generative IR models that have been proposed so far have achieved promising performance on large-scale document retrieval tasks [3, 45, 48]. However, in real-world scenarios, the IID assumption may not always be satisfied: the test distribution is usually unknown and possibly different from the training distribution. Put differently, high IID accuracy does not necessarily translate into out-of-distribution (OOD) robustness for document retrieval. Besides, pre-trained transformers, which usually serve as the backbone of existing generative IR models, may rely on spurious cues and annotation artifacts are less likely to include OOD examples [16]. So far, little is known about the OOD robustness of generative IR models.

In this work, we systematically study OOD robustness across various families of retrieval models, including generative, dense, and sparse retrieval models. In particular, we focus on comparing the robustness of generative IR models with that of the other models. We decompose OOD robustness into a model's generalization ability to (i) query variations, (ii) unseen query types, (iii) unseen tasks, and (iv) corpus expansion Each generalization ability perspective corresponds to a different OOD scenario. Based on this taxonomy, we design corresponding experiments and conduct empirical studies to analyze the robustness of several representative generative IR models against dense retrieval models.

For our experiments, we employ the comprehensive knowledge-intensive language tasks (KILT) benchmark [39], which comprises eleven datasets across five KILT tasks. With its *distinct tasks* and *multiple corpora* for several of its tasks, KILT is ideal for an analysis of OOD robustness. In the future we will also try to explore more general datasets and models for experimentation. In this work, following [3, 7], we consider the retrieval task of KILT, in which the model should retrieve a set of Wikipedia pages as evidence for the final prediction with respect to the input query.

Our experimental results reveal that, overall, generative IR models perform poorly in terms of OOD robustness. Different generative IR models display different types of generalizability performance in different OOD scenarios. As a result, there is considerable scope for future robustness improvements. With our findings, we draw attention to an understudied research area.

2 Related Work

2.1 Sparse and dense retrieval models

Sparse retrieval models build representations of queries and documents based on the bag-of-words (BoW) assumption [55], where each text is treated as a multiset of its words, ignoring grammar and word order [13, 41]. During the past decades, we have witnessed sparse retrieval models going through quick algorithmic shifts from early heuristic models [43], vector space models [43],

to probabilistic models [40, 41]. BM25 [42], as a representative of probabilistic models, is widely used for its efficiency while guaranteeing retrieval performance.

With the development of deep learning, many researchers have turned to dense retrieval models [19, 20, 56], which have been proven to be effective in capturing latent semantics and extracting effective features. Dense retrieval models typically adopt a bi-encoder architecture to encode queries and documents into low-dimension embeddings and use embedding similarities as estimated relevance scores for effective retrieval [13]. Karpukhin et al. [19] were pioneers in discovering that fine-tuning BERT to learn effective dense representations, called DPR, outperforms traditional retrieval methods like BM25. Subsequently, researchers began exploring various fine-tuning techniques to enhance dense retrieval models, such as mining hard negatives [51, 54], late interaction [20]. Recently, researchers have also investigated pre-training tasks for dense retrieval [12, 33]. Although these methods greatly improve the performance of dense retrieval models, they follow the same bi-encoder architecture represented by the DPR and usually come with considerable memory demands and computational overheads.

2.2 Generative IR models

Generative IR has recently garnered increasing interest [1, 2, 35, 48]. Generative IR retrieves documents by directly generating their identifiers based on the given query. It offers an end-to-end solution for document retrieval tasks [4, 35] and allows for better exploitation of the capabilities of large generative language models. For example, De Cao et al. [7] proposed an autoregressive entity retrieval model and Tay et al. [45] introduced a differentiable search index (DSI) and represent documents as atomic ids, naive string, or semantic strings. Chen et al. [3] proposed a pre-trained generative IR model called CorpusBrain to encode all information of the corpus within its parameters in a general way. Rather than using BART [23] directly as a generative IR model, CorpusBrain can capture relevance signals within documents, leading to promising retrieval performance in a wide range of knowledge-intensive language tasks. However, so far the robustness of generative IR models has been overlooked by the community.

2.3 Out-of-distribution in IR

OOD robustness refers to a model's ability to maintain performance when encountering data that differs from the distribution of the training data [15]. In real-world applications, retrieval models often face unseen data, highlighting the challenges of out-of-distribution robustness [25, 26, 31, 46, 49]. Current studies on OOD robustness in IR have their own limitations. For example, Wu et al. [49] only explored the OOD generalization performance of neural ranking models. Some work has been devoted to alleviating the poor performance of dense retrieval in the scenarios of query variants [6, 38, 44, 57] or zero/few-shot of corpus [24, 46, 53]. In this work, we focus on the OOD generalizability of generative IR models and compare them analytically with representative retrieval models from other families.

3 IID Settings for the Retrieval Problem

For a better understanding of the OOD setting for the retrieval problem, we first briefly introduce the IID setting for the retrieval problem.

Formally, given a dataset $\mathcal{D} = \{(q_i, D, Y_i)\}_{i=1}^{n}$, where q_i denotes a query, $D = \{d_1, d_2, \ldots, d_N\}$ represents the corpus, and $Y = \{r_1, r_2, \ldots, r_l\}$ indicates the corresponding relevance label of each document in D. A total order exists among the relevance labels such that $r_l \succ r_{l-1} \succ \cdots \succ r_1$, where $\succ$ denotes the order relation. Each query q_i is associated with a list of corresponding labels $\mathbf{y}_i = \{y_{i1}, y_{i2}, \ldots, y_{i,N}\}$, where N denotes the corpus size.

Traditionally, a retrieval model could be a term-based retrieval mode [40, 41] or a dense retrieval model [19, 33]. Recently, generative IR models have emerged as another paradigm [3, 7, 45]. Although the paradigm is different, these retrieval models have the same formal definition regarding the retrieval task. Without loss of generality, we use f to denote the retrieval model. We consider the retrieval model f learned on the dataset $\mathcal{D}$, which is drawn from the training distribution $\mathcal{G}$. For retrieval we employ the learned model f to generate a score for any query-document pair (q, d), reflecting the relevance degree of d given q. This set-up allows us to produce a permutation $\pi(q_t, D, f)$ according to predicted scores. Given an effectiveness evaluation metric M, retrieval models are typically evaluated by the average performance over the test queries under the IID setting, i.e.,

$$\mathbb{E}_{(q_t, D, \mathbf{y}_t) \sim \mathcal{G}} M(\pi(q_t, D, f), \mathbf{y}_t), \tag{1}$$

where q_t, D and $\mathbf{y}_t$ denote the query, the corpus and the label in the test set, respectively. Specifically, the test samples are supposed to be drawn from the same distribution as $\mathcal{G}$.

4 OOD Settings for the Retrieval Problem

In this work, we define the OOD robustness of retrieval models in four ways, i.e., in terms of query variations, unseen query types, unseen tasks, and corpus expansion. For query variations, the models are trained on the original dataset $\mathcal{D}$ and tested on the same dataset with query variations. For unseen query types and unseen tasks, the models are trained on an original dataset $\mathcal{D}$ and tested on a new dataset $\mathcal{D}'$ with the same task as $\mathcal{D}$ and on $\tilde{\mathcal{D}}$ with a task that is different from $\mathcal{D}$, respectively. For corpus expansion, the new dataset $\mathcal{D}^n$ that the models are tested on, is an expansion of the original dataset $\mathcal{D}$.

4.1 Query variations

The query variations refer to different expressions of the same information need. Therefore, a query and its variations usually correspond to the same related document. This query-level OOD aims to analyze the model's generalizability across different query variations within the dataset.

Formally, suppose that the examples q_t, D and $\mathbf{y}_t$ are drawn from the training distribution $\mathcal{G}$. We aim to evaluate the models' performance on the query OOD example. Specifically, the testing scenario of OOD generalizability on query variations is defined as

$$\mathbb{E}_{(q_t,D,\mathbf{y}_t)\sim\mathcal{G}} M(\pi(G(q_t), D, f), \mathbf{y}_t), \tag{2}$$

where $G(q_t)$ denotes the query variations generated by the generator G.

4.2 Unseen query types

The unseen query types scenario refers to unseen types of queries that are due to new types of information needs on the same task. Due to the query-specific provenance, query distributions differ between one query set to another, every though they focus on the same task. This query-type-level OOD aims to analyze the model's generalizability across different query types.

Formally, suppose that the new types of queries OOD examples q'_t and relevance label $\mathbf{y}'_t$ are drawn from the new distribution $\mathcal{G}'_Q$ and come from dataset $\mathcal{D}'$. The corpus of datasets $\mathcal{D}$ and $\mathcal{D}'$ are consistent as D. Specifically, the testing scenario of OOD generalizability on unseen query types is defined as

$$\mathbb{E}_{(q'_t,D,\mathbf{y}'_t)\sim\mathcal{G}'_Q} M(\pi(q'_t, D, f), \mathbf{y}'_t). \tag{3}$$

Note that the training dataset $\mathcal{D}$ and the testing dataset $\mathcal{D}'$ with unseen query types come from the same task.

4.3 Unseen tasks

The unseen tasks scenario refers to distribution shifts arising from task shifts. In practice, a retrieval model is usually trained to focus on a specific task and model a particular relevance pattern. Therefore, it is essential to evaluate how well a retrieval model, trained on datasets of a given task, can generalize to datasets of new tasks. This pair-level OOD aims to analyze a model's generalizability across different retrieval tasks.

Formally, suppose that the new task OOD examples $\tilde{q}_t$, $\tilde{D}$ and $\tilde{\mathbf{y}}_t$ are drawn from the new distribution $\tilde{\mathcal{G}_T}$ and come from dataset $\tilde{\mathcal{D}}$. Specifically, the testing scenario of OOD generalizability on unseen corpus is defined as

$$\mathbb{E}_{(\tilde{q}_t,\tilde{D},\tilde{\mathbf{y}}_t)\sim\tilde{\mathcal{G}_T}} M(\pi(\tilde{q}_t, \tilde{D}, f), \tilde{\mathbf{y}}_t). \tag{4}$$

Note that the training dataset $\mathcal{D}$ and the test dataset $\tilde{\mathcal{D}}$ belong to different tasks, respectively.

4.4 Corpus expansion

The corpus expansion scenario refers to the scenario for the trained IR model to maintain its retrieval performance under continuously arriving new documents.

Table 1: Statistics of datasets in the KILT benchmark. '-' denotes that the task does not provide ground-truth documents in the training set.

Task	Label	Dataset	Train. size	Dev. size
Dialogue	**WoW**	Wizard of Wikipedia [9]	63,734	3,054
Entity linking	**AY2**	AIDA CoNLL-YAGO [17]	18,395	4,784
	WnWi	WNED-WIKI [14]	-	3,396
	WnCw	WNED-CWEB [14]	-	5,599
Fact checking	**FEV**	FEVER [47]	104,966	10,444
Open domain QA	**NQ**	Natural Questions [21]	87,372	2,837
	HoPo	HotpotQA [52]	88,869	5,600
	TQA	TriviaQA [18]	61,844	5,359
	ELI5	ELI5 [11]	-	1,507
Slot filling	**T-REx**	T-REx [10]	2,284,168	5,000
	zsRE	Zero Shot RE [22]	147,909	3,724

In reality, an IR corpus may expand as new documents continuously enter the system. Along with this, queries related to these new documents will also emerge. Therefore, it is important to evaluate the model's adaptability to these unseen documents. This pair-level OOD analysis is aimed at assessing a model's ability to generalize to an expanding corpus.

Formally, suppose that the corpus update examples q_t^n, D^n and $\mathbf{y}_t^n$ come from corpus expansion $\mathcal{D}^n$. Specifically, the testing scenario of OOD generalizability on corpus expansion is defined as

$$\mathbb{E}_{(q_t, D, \mathbf{y}_t) \sim \mathcal{G}_T} M(\pi(q_t^n, D^n, f), \mathbf{y}_t^n), \tag{5}$$

Note that the test dataset $\mathcal{D}^n$ is an expansion of the training dataset $\mathcal{D}$.

5 Experimental Setup

We introduce the experimental setup for analyzing OOD robustness.

5.1 Datasets

For four OOD settings, we construct four benchmark datasets based on the KILT benchmark [39] (see Table 1). Due to the submission frequency limits of the online leaderboard, we used the performance on the dev set to evaluate model performance. In the following, we describe the details of the constructed datasets for evaluating the OOD generalizability on query variation, unseen query types, unseen tasks, and corpus expansion, respectively.

– **Dataset for query variations**. We use the queries in Fever (FEV) and Natural Questions (NQ) to generate their variations, as all of the retrieval

Table 2: Synthetic queries using variation generators.

Original query	who wrote most of the declaration of independence
Misspelling	who wr**eit** most of the declaration of independence
Naturality	~~**who**~~ wrote most ~~**of the**~~ declaration ~~**of**~~ independence
Order	who **declaration** most of the **wrote** of independence
Paraphrasing	who **authored** most of the declaration of independence

models perform relatively well on these datasets. Four generation strategies are considered [38] to perturb input queries, including (1) **Misspelling** for randomly substituting existing characters; (2) **Naturality** for removing all stop words; (3) **Order** for randomly exchanging positions of two words; and (4) **Paraphrasing** for replacing non-stop words according to the similarity of counter-fitted word embeddings [36]. Examples of the generated query variations are listed in Table 2.

- **Dataset for unseen query types**. We use the datasets under the open-domain QA task which covers the largest number of datasets in the KILT. There are three full datasets in open domain QA, i.e., NQ, HoPo, TQA. These datasets contain different topics and provenances, i.e., web search queries [21], multi-hop questions [52], and trivia questions [18].
- **Dataset for unseen tasks**. We use 5 tasks from the KILT benchmark, namely, dialogue (Dial.), entity linking (EL), fact checking (FC), open domain question answering (QA), and slot filling (SF). For each task in the KILT, we mix every training and test set of all datasets under each task separately to create a new dataset for that task.
- **Dataset for corpus expansion**. To mimic corpus expansion, we randomly sample 60% documents from the whole Wikipedia pages to serve as the initial corpus D_0 and leave the other 40% Wikipedia pages as the incremental corpus D_1. To construct the downstream KILT training set corresponding to D_0, We filter the original KILT training set by retaining only those query-document pairs where all relevant articles in the corresponding provenance exclusively belong to D_0. Similarly, to construct the test set Q_0 and Q_1 corresponding to D_0 and D_1, we first filter the original dev set by retaining only those query-document pairs where all relevant articles in the corresponding provenance exclusively belong to D_0, and then construct the filtered and remaining dataset as Q_0 and Q_1 respectively.

5.2 Retrieval models

We use representative samples of models from different families:

- **BM25** [42] is a representative sparse retrieval model that estimates the relevance based on term frequency, document length, and document frequency.
- **DPR** [19] is a representative dense retrieval model that uses dual-encoder architecture and is trained with in-batch negatives and a few hard negatives selected with BM25.

Table 3: R-precision (%) for the page-level retrieval task on the KILT dev data.

	Dial.	EL	FC	Open Domain QA			Slot Filling		
Model	**WoW**	**AY2**	**FEV**	**NQ**	**HoPo**	**TQA**	**T-REx**	**zxRE**	**Avg.**
BM25	27.5	3.5	50.1	25.8	44.0	29.4	58.6	66.4	38.2
DPR	25.2	2.1	52.9	53.9	26.1	42.8	13.5	28.4	30.6
BART	50.7	90.1	79.6	48.9	41.6	64.4	74.4	94.3	68.0
CorpusBrain	**55.0**	**90.7**	**81.4**	**57.6**	**50.7**	**70.9**	**75.7**	**97.6**	**72.5**

– **BART** [23] is a Seq2Seq model applicable for sequence generation tasks. Following [3, 7], we extract the query-title pairs from each dataset and fine-tune the BART for generative retrieval.
– **CorpusBrain** [3] is a pre-trained generative IR model for knowledge-intensive language tasks. We fine-tune CorpusBrain on every specific downstream KILT task.

5.3 Evaluation

To measure the OOD generalizability of the retrieval models, following [49], we use $\mathbf{DR}_{OOD}$ (%) to evaluate the drop rate between the retrieval performance P_{OOD} under the OOD setting and the retrieval performance P_{IID} under the IID setting, defined as,

$$DR_{OOD} = \frac{P_{OOD} - P_{IID}}{P_{IID}}, \tag{6}$$

where P_{IID} denotes the retrieval performance of the model trained on the training set corresponding to the test set. And P_{OOD} denotes the retrieval performance of the model trained on the training set that is out-of-distribution for the test set. The ranking model would be more robust with a higher DR_{OOD}.

The effectiveness metric for evaluating the retrieval performance in KILT is usually defined as **R-precision** (%), which is suggested in the official instructions and widely used in previous works on KILT [2, 3, 7]. R-precision is calculated as $\frac{r}{R}$, where R is the number of Wikipedia pages inside each provenance set and r is the number of relevant pages among the top-R retrieved pages.

6 Results

We examine the empirical results in the IID setting and the three OOD settings sequentially: query variations, unseen query types, unseen tasks, and corpus expansion.

6.1 Overall IID results

We compare the selected retrieval models on the KILT benchmark. From Table 3, we can observe that the generative IR models significantly outperform sparse

Table 4: R-precision / DR_{OOD} for query variations on the FEV and NQ dev data. Significant performance degradation with respect to the corresponding IID setting is denoted as '−' ($p\text{-}value \leq 0.05$).

Model	Original	Misspelling	Naturality	Order	Paraphrasing
			FEV		
BM25	50.1	**31.8**/-36.5	**49.2**/-0.02	22.3/ 0	39.5/-21.2
DPR	52.9	24.1/-54.4	32.4/-38.8	22.3/-57.8	34.8/-34.2
BART	79.6	20.7/-74.0	38.3/-51.9	22.1/-72.2	34.7/-56.4
CorpusBrain	**81.4**	26.0/-68.0	41.8/–48.6	**27.7**/-66.0	**40.6**/-50.1
			NQ		
BM25	25.8	20.5/-20.5	25.4/-0.02	31.0/ 0	22.1/-14.3
DPR	53.9	25.6/-52.5	31.8/-41.0	31.0/-42.5	44.6/-17.3
BART	48.9	26.2/-46.4	39.1/-20.0	32.8/-32.9	43.4/-11.2
CorpusBrain	**57.6**	**28.1**/-51.2	**39.2**/-31.9	**36.1**/-37.3	**50.1**/-13.0

and dense retrieval models like BM25 and DPR across all the datasets, indicating that combining the retrieval components into a unified model benefits effective corpus indexing. CorpusBrain consistently outperforms BART on all five tasks, demonstrating that the adequately well-designed pre-training tasks for generative retrieval contribute to improving document understanding for generative IR models.

6.2 Analysis of OOD generalizability on query variations

Firstly, from Table 4, we provide a comprehensive performance analysis of all the retrieval models. We can observe that some types of query variants, such as naturality and order, have little impact on BM25. This is because BM25 uses bags of words to model documents and is insensitive to changes in word order and stop words. Beyond that, there is a significant effectiveness drop for query variations in all dense and generative retrieval models. The results indicate that both dense retrieval models, as well as generative IR models, are not robust to query variations, which complements the findings from previous work [38].

When we compare the generalizability of the dense and generative IR models, we find that the generative IR models perform particularly poorly on **Misspelling** and **Order**. One possible explanation would be that the generative IR models generate document identifiers autoregressively based on the query, so query quality and word order greatly impact the generation effect. When we look at the performance of generative IR models, we can find that the CorpusBrain has better R-precision than BART, indicating that pre-training tasks tailored for generative retrieval help the model adapt better to query variations.

Table 5: R-precision/DR_{OOD} for unseen query types on the open domain QA dataset of KILT.

Model	Training	Testing		
		NQ	**HoPo**	**TQA**
BM25	NQ	25.8	-	-
	HoPo	-	44.0	-
	TQA	-	-	29.4
DPR	NQ	53.9	23.1/-11.5	29.2/-31.8
	HoPo	41.2/-23.6	26.1	26.3/-38.6
	TQA	42.3/-21.5	21.8/-16.5	42.8
BART	NQ	48.9	36.4/-12.5	50.7/-21.3
	HoPo	18.8/-61.6	41.6	46.8/-27.3
	TQA	26.5/-45.8	35.2/-15.4	64.4
CorpusBrain	NQ	**57.6**	47.0/ -7.3	52.7/-25.7
	HoPo	33.4/-42.0	**50.7**	48.6/-31.5
	TQA	32.9/-42.9	44.7/-11.8	**70.9**

6.3 Analysis of OOD generalizability on unseen query types

The results of OOD generalizability on unseen query types are shown in Table 5. Note that BM25 does not rely on the training set, so its test results remain consistent across datasets. When we look at the overall performance of all the dense and generative retrieval models, we can observe that as the shift of query types distributions, the performance of all models decreases significantly. For DPR, some of the query types in which it had an advantage are instead inferior to BM25 in OOD scenarios. This suggests that, even under the same task, neural retrieval models face challenges of poor OOD generalizability. Consequently, it is important to consider the OOD performance for these unseen query types.

Comparing the performance of dense and generative IR models, we find that generative IR models exhibit worse generalizability on web search queries in the NQ dataset. It indicates that, in terms of generalizability performance on unseen query types, generative IR models behave differently and merit separate studies. Furthermore, we observe that CorpusBrain demonstrates better generalizability than BART on unseen query types. This could be attributed to the pre-training process of CorpusBrain, which effectively encodes relevant information for a given corpus to cope with potentially unknown queries, thereby enhancing its stability when encountering unseen query types.

6.4 Analysis of OOD generalizability on unseen tasks

Examining the overall performance of all retrieval models in Table 6, we observe that the generalizability defects for unseen tasks are common among models. In the entity linking (EL) task, the models' generalization performance drops significantly, likely due to the task's distinct format compared to the others.

Table 6: R-precision/DR_{OOD} for unseen tasks on the 5 KILT task-mixed datasets.

		Testing				
Model	**Training**	**Dial.**	**EL**	**FC**	**QA**	**SF**
BM25	Dial.	27.5	-	-	-	-
	EL	-	3.5	-	-	-
	FC	-	-	50.1	-	-
	QA	-	-	-	34.6	-
	SF	-	-	-	-	61.9
DPR	Dial.	25.5	0.7/-66.7	48.2/ -8.9	25.1/-30.1	10.3/-46.6
	EL	13.6/-46.7	2.1	46.8/-11.5	15.9/-55.7	13.6/-29.5
	FC	24.0/ -5.9	0.8/-61.9	52.9	22.3/-37.9	15.2/-21.1
	QA	22.5/-11.8	0.6/-71.4	50.8/ -4.0	35.9	12.1/-37.3
	SF	10.0/-60.8	0.2/-90.5	45.4/-14.2	20.2/-43.7	19.3
BART	Dial.	49.7	8.6/-90.5	71.8/ -9.8	40.2/-31.8	69.2/-17.5
	EL	21.5/-56.7	90.1	68.6/-13.8	24.6/-58.3	77.4/ -7.7
	FC	48.1/ -3.2	10.0/-88.9	79.6	41.5/-29.7	81.1/ -3.3
	QA	45.0/ -9.5	9.8/-89.1	76.4/ -4.0	59.0	77.6/ -7.5
	SF	17.1/-65.6	3.7/-95.9	65.6/-17.6	36.3/-38.5	83.9
CorpusBrain	Dial.	**58.0**	6.9/-92.4	74.1/ -9.0	49.3/-18.8	77.7/ -7.8
	EL	33.0/-43.1	**90.7**	68.6/-15.7	38.4/-36.7	62.7/-25.6
	FC	46.8/-19.3	9.3/-89.7	**81.4**	48.9/-19.4	82.2/ -2.5
	QA	46.3/-20.1	8.1/-91.1	79.2/ -2.7	**60.7**	78.7/ -6.6
	SF	25.3/-56.4	4.9/-94.6	68.1/-16.3	43.7/-28.0	**84.3**

DPR lags behind BM25 almost across the board when faced with unseen tasks. The reason may be the large differences in data distribution across tasks. The semantic representations that DPR learns by learning from the original task are empirical and difficult to flexibly migrate to the new task. While dense retrieval models have excellent performance, there are situations where traditional sparse retrieval models are rather more to be relied upon.

When we observe the performance between dense and generative IR models, we find that, in general, generative IR models have higher DR_{OOD} on slot-filling (SF) task. This could be because the format of this downstream task aligns with the pre-training tasks of the backbone generative models. Comparing BART and CorpusBrain from the generative IR models, we observe that CorpusBrain outperforms BART in most (13 out of 20) unseen task scenarios. This may be attributed to the pre-training tasks of CorpusBrain. CorpusBrain includes three tasks: Inner Sentence Selection (ISS), Lead Paragraph Selection (LPS), and Hyperlink Identifier Prediction (HIP). ISS models the semantic granularity differences between queries and documents in various retrieval requirements, helping to bridge the gap between different downstream tasks. This finding is consistent with the original analysis of CorpusBrain [3].

Table 7: R-precision/DR_{OOD} for corpus expansion on the 5 KILT task-specific datasets.

Model	Session	Dial.	EL	FC	QA	SF
BM25	$\mathbf{D}_0$	26.0	2.6	46.5	40.2	55.7
	$\mathbf{D}_1$	20.5/-21.2	1.8/-30.8	37.8/ -6.0	28.2/-29.9	46.7/-16.2
DPR	$\mathbf{D}_0$	49.2	2.3	73.2	46.5	40.1
	$\mathbf{D}_1$	28.7/-41.7	1.6/-30.4	65.7/-10.2	41.8/-10.1	35.0/-12.7
BART	$\mathbf{D}_0$	42.7	63.0	74.5	22.7	63.6
	$\mathbf{D}_1$	47.0/ 10.1	57.2/ -9.2	70.7/ -5.1	20.9/ -7.9	47.0/-26.1
CorpusBrain	$\mathbf{D}_0$	36.4	64.0	78.8	41.2	81.2
	$\mathbf{D}_1$	15.1/-58.5	43.5/-32.0	56.7/-28.0	28.3/-31.3	74.7/ -8.0

6.5 Analysis of OOD generalizability on corpus expansion

The result of OOD generalizability on corpus expansion is shown in Table 7. From the result, we can observe that, BM25 has an average ability to maintain retrieval performance. When faced with incremental documents entering the corpus, the performance degradation of DPR is not significant. The possible reason for this is that under the same task, the newly arrived document belongs to the same topic as the old one, and DPR can build the complete semantic representation space from the original training.

For the generative IR model, both BART and CorpusBrain perform significantly better in corpus expansion than dense retrieval models like DPR. Even with a lower DR_{OOD} than BM25 and DPR, the overall performance of the generative IR model is still higher than that of them. The reason for this is that generative retrieval uses a prefix tree to store indexes, and unseen indexes will still be distributed in the neighborhood of the indexes they are related to. That is, generative IR models can extensively probe for relevant documents in the corpus through beam search, which underpins their generalization capabilities.

7 Conclusion

In this paper, we have analyzed the out-of-distribution robustness of several representative generative and dense retrieval models on the KILT benchmark. Specifically, we have proposed four perspectives to define out-of-distribution robustness. Our results exposed significant vulnerabilities in OOD robustness of generative IR models.

We believe that the understanding of different forms of retrieval models can open up ideas from a robustness perspective. As we observed, the dense retrieval model and the generative retrieval model perform differently for different OOD scenarios. While there is some prior work that relates generative and dense retrieval [37, 50], the robustness perspective on their connection is missing – what can we learn from their relative strengths and weaknesses to develop more robust retrieval models?

Concerning limitations presented here, we chose CorpusBrain and BART, which perform well on KILT, as representatives of generative IR models. In future work, we will introduce more generative IR models with more datasets to further explore the OOD robustness. Due to inheriting the vulnerabilities of neural network models, neural IR models are also susceptible to being deceived by out-of-distribution adversarial examples [27, 28, 29, 30, 32]. Future work should consider introducing new web search datasets into the benchmark to simulate more broader and potentially even more challenging OOD environments like adversarial attacks. Our work highlights the need to create benchmarks that include various OOD perspectives to better understand the generative IR models' robustness.

Finally, we will consider different docid forms of generative IR models to explore the differences in robustness performance between generative IR models.

Acknowledgements

This work was funded by the National Natural Science Foundation of China (NSFC) under Grants No. 62472408, the Strategic Priority Research Program of the CAS under Grants No. XDB0680102, the National Key Research and Development Program of China under Grants No. 2023YFA1011602, the Lenovo-CAS Joint Lab Youth Scientist Project, and the project under Grants No. JCKY2022130C039. This work was also (partially) funded by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union's Horizon Europe program under grant agreement No. 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Bibliography

[1] Bevilacqua, M., Ottaviano, G., Lewis, P., Yih, S., Riedel, S., Petroni, F.: Autoregressive search engines: Generating substrings as document identifiers. Advances in Neural Information Processing Systems **35**, 31668–31683 (2022)

[2] Chen, J., Zhang, R., Guo, J., Fan, Y., Cheng, X.: Gere: Generative evidence retrieval for fact verification. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2184–2189 (2022)

[3] Chen, J., Zhang, R., Guo, J., Liu, Y., Fan, Y., Cheng, X.: Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 191–200 (2022)

[4] Chen, J., Zhang, R., Guo, J., de Rijke, M., Liu, Y., Fan, Y., Cheng, X.: A unified generative retriever for knowledge-intensive language tasks via prompt learning. In: SIGIR, pp. 1448–1457 (2023)

[5] Chen, R.C., Gallagher, L., Blanco, R., Culpepper, J.S.: Efficient cost-aware cascade ranking in multi-stage retrieval. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 445–454 (2017)

[6] Chen, X., Luo, J., He, B., Sun, L., Sun, Y.: Towards robust dense retrieval via local ranking alignment. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, pp. 1980–1986 (2022)

[7] De Cao, N., Izacard, G., Riedel, S., Petroni, F.: Autoregressive entity retrieval. In: International Conference on Learning Representations (2020)

[8] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)

[9] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., Weston, J.: Wizard of wikipedia: Knowledge-powered conversational agents. In: International Conference on Learning Representations (2018)

[10] Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Laforest, F., Simperl, E.: T-rex: A large scale alignment of natural language with knowledge base triples. In: LREC 2018 (2018)

[11] Fan, A., Jernite, Y., Perez, E., Grangier, D., Weston, J., Auli, M.: Eli5: Long form question answering. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3558–3567 (2019)

[12] Gao, L., Callan, J.: Condenser: a pre-training architecture for dense retrieval. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021)

[13] Guo, J., Cai, Y., Fan, Y., Sun, F., Zhang, R., Cheng, X.: Semantic models for the first-stage retrieval: A comprehensive review. ACM Transactions on Information Systems **40**(4), 1–42 (2022)

[14] Guo, Z., Barbosa, D.: Robust named entity disambiguation with random walks. Semantic Web **9**(4), 459–479 (2018)

[15] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)

[16] Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., Song, D.: Pretrained transformers improve out-of-distribution robustness. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2744–2751 (2020)

[17] Hoffart, J., Yosef, M.A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., Weikum, G.: Robust disambiguation of named entities in text. In: Proceedings of the 2011 conference on empirical methods in natural language processing, pp. 782–792 (2011)

[18] Joshi, M., Choi, E., Weld, D., Zettlemoyer, L.: TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In: ACL, pp. 1601–1611, Association for Computational Linguistics, Vancouver, Canada (Jul 2017)

[19] Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., tau Yih, W.: Dense passage retrieval for open-domain question answering. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781, Association for Computational Linguistics (2020)

[20] Khattab, O., Zaharia, M.A.: ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval pp. 39–48 (2020)

[21] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.W., Dai, A.M., Uszkoreit, J., Le, Q., Petrov, S.: Natural questions: A benchmark for question answering research. Transactions of the Association for Computational Linguistics **7**, 453–466 (2019)

[22] Levy, O., Seo, M., Choi, E., Zettlemoyer, L.: Zero-shot relation extraction via reading comprehension. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pp. 333–342 (2017)

[23] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7871–7880 (2020)

[24] Liang, D., Xu, P., Shakeri, S., dos Santos, C.N., Nallapati, R., Huang, Z., Xiang, B.: Embedding-based zero-shot retrieval through query generation. arXiv preprint arXiv:2009.10270 (2020)

[25] Liu, Y.A., Zhang, R., Guo, J., de Rijke, M.: Robust information retrieval. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3009–3012 (2024)
[26] Liu, Y.A., Zhang, R., Guo, J., de Rijke, M.: Robust information retrieval. In: Proceedings of the 18th ACM International Conference on Web Search and Data Mining (2025)
[27] Liu, Y.A., Zhang, R., Guo, J., de Rijke, M., Chen, W., Fan, Y., Cheng, X.: Black-box adversarial attacks against dense retrieval models: A multi-view contrastive learning method. In: CIKM, p. 1647–1656 (2023)
[28] Liu, Y.A., Zhang, R., Guo, J., de Rijke, M., Chen, W., Fan, Y., Cheng, X.: Topic-oriented adversarial attacks against black-box neural ranking models. In: SIGIR, p. 1700–1709 (2023)
[29] Liu, Y.A., Zhang, R., Guo, J., de Rijke, M., Cheng, X.: Attack-in-the-chain: Bootstrapping large language models for attacks against black-box neural ranking models. In: Proceedings of the AAAI Conference on Artificial Intelligence (2025)
[30] Liu, Y.A., Zhang, R., Guo, J., de Rijke, M., Fan, Y., Cheng, X.: Multi-granular adversarial attacks against black-box neural ranking models. In: SIGIR (2024)
[31] Liu, Y.A., Zhang, R., Guo, J., de Rijke, M., Fan, Y., Cheng, X.: Robust neural information retrieval: An adversarial and out-of-distribution perspective. arXiv preprint arXiv:2407.06992 (2024)
[32] Liu, Y.A., Zhang, R., Zhang, M., Chen, W., de Rijke, M., Guo, J., Cheng, X.: Perturbation-invariant adversarial training for neural ranking models: Improving the effectiveness-robustness trade-off. In: AAAI, vol. 38 (2024)
[33] Ma, X., Zhang, R., Guo, J., Fan, Y., Cheng, X.: A contrastive pre-training approach to discriminative autoencoder for dense retrieval. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 4314–4318 (2022)
[34] Matveeva, I., Burges, C., Burkard, T., Laucius, A., Wong, L.: High accuracy retrieval with multiple nested ranker. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 437–444 (2006)
[35] Metzler, D., Tay, Y., Bahri, D., Najork, M.: Rethinking search: making domain experts out of dilettantes. ACM SIGIR Forum **55**(1), 1–27 (2021)
[36] Mrkšić, N., Séaghdha, D.Ó., Thomson, B., Gasic, M., Rojas-Barahona, L.M., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Counter-fitting word vectors to linguistic constraints. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 142–148 (2016)
[37] Nguyen, T., Yates, A.: Generative retrieval as dense retrieval. arXiv preprint arXiv:2306.11397 (2023)
[38] Penha, G., Câmara, A., Hauff, C.: 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)

[39] Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., Riedel, S.: KILT: A benchmark for knowledge intensive language tasks. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2523–2544, Association for Computational Linguistics, Online (Jun 2021)

[40] Ponte, J., Croft, W.B.: A language modeling approach to information retrieval. In: SIGIR, pp. 275–281 (1998)

[41] Robertson, S., Walker, S.: Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In: Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 232–241 (1994)

[42] Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009)

[43] Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM **18**(11), 613–620 (1975)

[44] Sidiropoulos, G., Kanoulas, E.: Analysing the robustness of dual encoders for dense retrieval against misspellings. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, p. 2132–2136 (2022)

[45] Tay, Y., Tran, V.Q., Dehghani, M., Ni, J., Bahri, D., Mehta, H., Qin, Z., Hui, K., Zhao, Z., Gupta, J., et al.: Transformer memory as a differentiable search index. arXiv preprint arXiv:2202.06991 (2022)

[46] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2021)

[47] Thorne, J., Vlachos, A., Christodoulopoulos, C., Mittal, A.: Fever: A large-scale dataset for fact extraction and verification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 809–819 (2018)

[48] Wang, Y., Hou, Y., Wang, H., Miao, Z., Wu, S., Sun, H., Chen, Q., Xia, Y., Chi, C., Zhao, G., et al.: A neural corpus indexer for document retrieval. arXiv preprint arXiv:2206.02743 (2022)

[49] Wu, C., Zhang, R., Guo, J., Fan, Y., Cheng, X.: Are neural ranking models robust? ACM Transactions on Information Systems **41**(2), 1–36 (2022)

[50] Wu, S., Wei, W., Zhang, M., Chen, Z., Ma, J., Ren, Z., de Rijke, M., Ren, P.: Generative retrieval as multi-vector dense retrieval. In: SIGIR 2024: 47th international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1828–1838, ACM (July 2024)

[51] Xiong, L., Xiong, C., Li, Y., Tang, K.F., Liu, J., Bennett, P., Ahmed, J., Overwijk, A.: Approximate nearest neighbor negative contrastive learning for dense text retrieval. arXiv preprint arXiv:2007.00808 (2021)

[52] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., Manning, C.D.: HotpotQA: A dataset for diverse, explainable multi-hop question answering. In: EMNLP, pp. 2369–2380, Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018)
[53] Yu, Y., Xiong, C., Sun, S., Zhang, C., Overwijk, A.: COCO-DR: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. arXiv preprint arXiv:2210.15212 (2022)
[54] Zhan, J., Mao, J., Liu, Y., Guo, J., Zhang, M., Ma, S.: Optimizing dense retrieval model training with hard negatives. Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval pp. 1503–1512 (2021)
[55] Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: A statistical framework. International Journal of Machine Learning and Cybernetics **1**, 43–52 (2010)
[56] Zhao, W.X., Liu, J., Ren, R., Wen, J.R.: Dense text retrieval based on pretrained language models: A survey. ACM Transactions on Information Systems **42**(4), 1–60 (2024)
[57] Zhuang, S., Zuccon, G.: Dealing with typos for BERT-based passage retrieval and ranking. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2836–2842 (2021)