Challenges in Generalization in Open Domain Question Answering

Anonymous ACL submission

Abstract

Recent work on Open Domain Question Answering has shown that there is a large discrepancy in model performance between *novel* test questions and those that largely overlap with training questions. However, it is as of yet unclear which aspects of novel questions that make them challenging. Drawing upon studies on systematic generalization, we introduce and annotate questions according to three categories that measure different levels and kinds of generalization: *training set overlap*, *compositional generalization (comp-gen)*, and *novel entity generalization (novel-entity)*. When evaluating six popular parametric and non-parametric models, we find that for the established Natural Questions and TriviaQA datasets, even the strongest model performance for comp-gen/novel-entity is 13.1/5.4% and 9.6/1.5% lower compared to that for the full test set – indicating the challenge posed by these types of questions. Furthermore, we show that whilst non-parametric models can handle questions containing novel entities, they struggle with those requiring compositional generalization. Through thorough analysis we find that key question difficulty factors are: cascading errors from the retrieval component, frequency of question pattern, and frequency of the entity.

1 Introduction

Over the last few years we have seen model innovations improving on standard natural language processing (NLP) benchmarks across the board (Devlin et al., 2019; Raffel et al., 2019; Lewis et al., 2020a). However, it is still clear that we are yet to obtain generalizable language understanding, as recent work has found that adversarial (Jia and Liang, 2017; Mudrakarta et al., 2018; Belinkov and Bisk, 2018) and out-of-distribution samples (Talmor and Berant, 2019; Elsahar and Gallé, 2019; McCoy et al., 2020) remain challenging for existing models across numerous tasks.

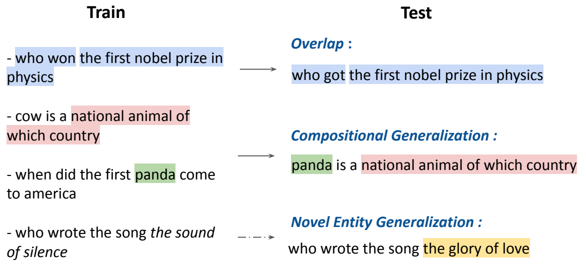
Train
Test
- who won the first nobel prize in physics
Overlap :
who got the first nobel prize in physics
- cow is a national animal of which country
Compositional Generalization :
- when did the first panda come to america
panda is a national animal of which country
- who wrote the song the sound of silence
Novel Entity Generalization :
who wrote the song the glory of love

Figure 1: Questions categorized according to their relation to the training set: 1) *Overlap*: there exists a paraphrase of the question in the training set. 2) *Compositional*: all individual facts and the structure of the question has been observed across several questions in the training set – but not the given composition. 3) *Novel-entity*: the question contains at least one entity (marked here with yellow) not present in the training set.

Open-domain question answering (ODQA), which aims to answer factoid questions without any given context, is a task that has been receiving increasing attention in the community (Chen et al., 2017; Lee et al., 2019; Karpukhin et al., 2020; Izacard and Grave, 2020; Min et al., 2021). However, recent work has shown that there is a large discrepancy in model performance between questions and answers observed at train time and *novel* questions and answers – even if they are derived from the same distribution (Lewis et al., 2020c). This raises the question: "What are the aspects of these novel questions that make generalization challenging?" This is the question which we seek to explore in this paper.

In work on systematic generalization (Bahdanau et al., 2018; Lake and Baroni, 2018; Ruis et al., 2020), it is argued that even though a model has only observed a very small subset of all possible combinations of facts at training time, a good model should be able to generalize to all possible combinations of facts at test time. We draw upon these ideas to study generalization for ODQA and define the following three categories to aid our investigation: *training set overlap*, *compositional generalization*, and *novel-entity generalization*. See Figure 1 for category definitions and examples. Our breakdown into three categories is motivated by how they capture different levels of generalization: *overlap* requiring no generalization beyond recognizing paraphrases, *comp-gen* requiring generalization to novel compositions of previously observed entities and structures, and *novel-entity* requiring generalization to entities not present in the training set. It is worth noting here that we explicitly study in-distribution generalization rather than out-of-distribution generalization (such as cross-domain generalization (Fisch et al., 2019)), as we will later demonstrate that even in-distribution generalization poses a major challenge for existing approaches.

We decompose and manually annotate three previously introduced ODQA datasets (Natural Questions (Lee et al., 2019), TriviaQA (Joshi et al., 2017), and WebQuestions (Berant et al., 2013)). Following this, we evaluate six recently proposed non-parametric and parametric ODQA models and analyze their performance, using both aggregate metrics and a breakdown according to our proposed categories. Non-parametric and parametric models differ in their access to information: the former has no access to any external context or knowledge, whereas the latter is provided relevant information alongside the question (Roberts et al., 2020). What we find is that both the non-parametric and parametric models perform, as expected, strongly on the *overlap* subset. However, both perform substantially worse on both *comp-gen* and *novel-entity*, but the parametric models perform significantly worse than their non-parametric counterparts.

Among the non-parametric models, FiD (Izacard and Grave, 2020) greatly outperforms the other models in its category. However, it still performs substantially worse for subsets that require generalization and we thus use it as a case study as to why non-parametric models underperform on these subsets. Firstly, we note that its retrieval component (Karpukhin et al., 2020, DPR) – a core part of any non-parametric model – finds it more challenging to retrieve relevant passages for *comp-gen* and *novel-entity* questions. To quantify the impact of this issue, if DPR could retrieve passages with the same accuracy for these two subsets as for *overlap*, FiD's performance for Natural Questions would increase by $\sim 4\%$.

Another potential source of difficulty could be the question structure itself and as a byproduct of our decomposition approach we are able to derive a high-level *question pattern* for each question – for example, for the question "who got the first nobel prize in physics" we derive the pattern "who got [entity]". When we examine performance across different patterns, we find a strong positive correlation between the pattern training frequency and accuracy. To quantify this issue, if FID was to perform equally well for all question patterns regardless of frequency, its performance for Natural Questions would increase by $\sim 11\%$.

Lastly, and perhaps most surprisingly, we find that the frequency of entities mentioned in a question is strongly *negatively* correlated with accuracy. Based on a thorough manual inspection, we find that this is most prominent for the comp-gen subset and likely due to highly frequent entities leading to a very large number of superficially relevant distractor passages being retrieved, which in turn leads to a performance degradation for the reader component. To quantify this issue, if FiD performed equally well across all entities regardless of their frequency in the training set, its performance for Natural Questions would increase by $\sim 4\%$.

To conclude, our key contributions are as follows:

1. We provide the first detailed study on generalization for ODQA, based on categories that measure different levels and kinds of generalization, that we use to annotate three previously proposed ODQA datasets.

2. We show that non-parametric models struggle to perform compositional generalization, whereas they handle novel question entities comparatively well.

3. We demonstrate and quantify that the key factors that impact generalization performance are: the retrieval component, frequency of question pattern, and frequency of the entity.

2 Dataset Construction

In this section, we describe how we process and annotate ODQA datasets to enable us to investigate generalization.

2.1 Question Decompostition

To study the compositional and novel-entity generalization of questions, we follow Keysers et al.

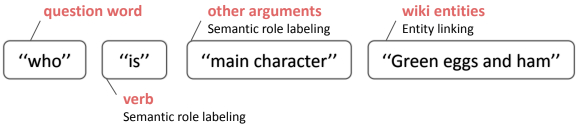

Figure 2: Example decomposition for the question *"Who is the main character in Green eggs and ham?"*

Group	Natural Questions	WebQ	TriviaQA
Overlap	837	501	458
Comp-gen	1,105	512	475
Novel-entity	597	640	456

Table 1: Number of questions for each generalization subset for the three datasets' test sets

(2019) and propose to view each question as being composed of primitive elements (atoms). Consider the question *"Who got the first Nobel Prize in Physics?"*. The atoms intuitively corresponds to the modifier or adjunct of the predicate "who", predicate "got" and the entity "first nobel prize in physics". The combination of these atoms cover the main semantics of the question.

The way we measure generalization necessarily depends on how we break down the questions into their atoms. Following manual analysis of questions from three popular ODQA datasets, we developed the following decomposition strategy to obtain atoms which cover all the desired question semantics. These are: question words, verbs, Wikipedia named entities (*wiki_entities*), and finally, other arguments (*other_args*) which correspond to other relevant aspects of the question. We explicitly extract wiki_entities since they leverage crucial semantics in factoid questions and other_args define essential details surrounding wiki_entities.

In order to automatically decompose questions, we first use an off-the-shelf semantic role labeling (SRL) model (Shi and Lin, 2019) to produce predicate-argument structures for each question. This provides us with the verb (i.e. the predicate), and semantic arguments. The question word is trivially obtained by identifying WH-words. We apply an off-the-shelf entity linking model (Li et al., 2020) to obtain the wiki_entities in the question. Finally, other_args are the SRL arguments which remain after we filter out arguments corresponding to wiki_entities. An example question decomposition is illustrated in Figure 2. More details are included in Appendix A.1.

2.2 Generalization Category Definitions

Based on the question decomposition, we define three generalization categories for ODQA datasets. We denote S_q as the set of the decomposed atoms of question q and C_Q as the complete set of decomposed atoms for all the questions in dataset Q. Our category subsets are then defined as:

- $Q_{\text{overlap}} \triangleq \{q \in Q_{\text{test}} \mid \exists\, q' \in Q_{\text{train}}, S_q \subseteq S_{q'}\}$
- $Q_{\text{comp\_gen}} \triangleq \{q \in Q_{\text{test}} \mid \exists\, q'_1, q'_2, ..., q'_k, S_q \subset C_{\text{train}},\ S_q \subseteq \bigcup_{i=1}^{k} S_{q'_i},\ S_q \nsubseteq S_{q'_i}\}$
- $Q_{\text{novel\_entity}} \triangleq \{q \in Q_{\text{test}} \mid \exists\, s \in S_q, s \notin C_{\text{train}}\}$

For overlap test question, there exists a train question where they have the same decomposed atoms or are subset of them; for comp_gen test question, its decomposed atoms are fully covered by the training set (a subset of the union of multiple train questions atoms), but not in one particular train question; and for novel-entity test question, there exist wiki_entities not present in the training set.

2.3 Question Categorization and Human Verification

With the decomposed atoms for all questions, we first categorize the test questions into overlap, comp-gen, and novel-entity categories based on the definitions of each generalization category. We optimize the selection criteria to cover as many eligible candidates for each category as possible. Further details can be found in Appendix A.2.

As our test set subsets are obtained automatically, we need to perform manual human verification to ensure that they are of high enough quality to draw empirical conclusions. To do this, we employ four expert annotators and use the following annotation process for each of the respective categories. *Overlap:* Annotators are shown q_{test} and the training questions with the highest degree of character-level overlap. If any of these questions are a paraphrase of q, the annotator will mark q_{test} as an *overlap* question. *Comp-gen:* q_{test} is presented to the annotators along with the training questions with the highest degree of word overlap. Annotators then verify that the test question is truly a compositional generalization and not a paraphrase of any of the

Group	Test question	Paired train question for annotator	Label
Overlap	who got the first nobel prize in physics	who won the first nobel prize in physics	T
	whens the last time the patriots played the eagles	when did the philadelphia eagles last win the super bowl	F
Comp-gen	when is the next scandal episode coming out	when is next fairy tail episode coming out	T
	what is the corporate tax rate in great britain	what is the rate of corporation tax in uk	F
Novel-entity	who wrote the song the *glory of love*	who sang *guilty of love* in the first degree	T
	who sings *too much time on my hands* lyrics	who sings *i've got too much time on my hands*	F

Table 2: Example of questions from Natural Questions (see Appendix A.8 for examples from the other two datasets) for human verification and their respective annotated labels (T for True and F for False).

given training questions. *Novel-entity:* Annotators need to: 1) Verify that the wiki_entities identified by the entity-linking model are indeed wiki entities. 2) Verify that the entities in q_{test} are not present among a set of questions from the training set whose entities have a high degree of character-level overlap with the entities in q_{test}. Statistics for the annotated category subsets are summarized in Table 1, examples are shown in Table 2, and additional details covered in Appendix A.3.

3 Experiment

3.1 Datasets

We analyse three widely used ODQA datasets, each one is briefly introduced as follows:

Open Natural Questions (NQ) is an open-domain variant of Natural Questions (Kwiatkowski et al., 2019) introduced by Lee et al. (2019) and it consists of 79,168 train, 8,757 dev, and 3,610 test question answer pairs.

TriviaQA (Joshi et al., 2017) consists of questions and answers which were obtained by scraping trivia websites. We use the open domain splits (Lee et al., 2019) which contains 78,785 train, 8,837 dev, and 11,313 test question answer pairs. For our analyses, we randomly sampled and annotated 2,000 questions from the test set.

WebQuestions (Berant et al., 2013) consists of questions that were collected by performing a breadth-first search using the Google Suggest API and it contains 3,778 train and 2,032 test questions.

3.2 Baseline Models

Non-parametric Models mostly adopt a retrieve-and-read framework, retrieving relevant Wikipedia documents for the given question, and then produce the final answer conditioned on these documents. We consider two generative reader models: Retrieval-Augmented Generation (RAG, Lewis et al., 2020b), and Fusion-In-Decoder (FiD, Izacard and Grave, 2020). RAG combines a DPR (Karpukhin et al., 2020) dense retriever with a BART (Lewis et al., 2020a) generator, which are jointly fine-tuned end to end. FiD is a pipeline approach which uses DPR to retrieve a set of documents, the decoder attends over the concatenation of all encoded document representations to generate the final answer. As an extractive reader model we use the reader component from DPR (Karpukhin et al., 2020). It extracts answer span from the highest-scoring document ranked from a passage selection model. We also include RePAQ (Lewis et al., 2021), a QA-pair retriever which does *not* follow the retrieve-and-read paradigm. RePAQ is based on dense Maximum Inner Product Search (MIPS) paired with an ALBERT-based reranker. It retrieves QA-pairs from PAQ, a large resource of 65M automatically-generated QA-pairs, finding the most relevant QA-pair and returning this answer.

Parametric Models are directly trained with QA pairs without access to an external corpus and thus store the required knowledge in its entirety in the model parameters. For our analyses, we include a BART-large model (Lewis et al., 2020a) and a more powerful T5-11B model (Roberts et al., 2020). They are both trained with questions as input and output question-answer pairs.

3.3 Model Category Analysis

Table 3 shows the Exact Match scores for models on our test set splits.

Overlap questions can be memorized All models perform significantly higher on overlap questions. This is consistent with the findings of Lewis et al. (2020c), who also show that models perform significantly better on test questions which have been seen during training. Parametric models with more parameters are the most effective at rote-memorizing training questions, and T5-11B+SSM even outperforms the non-parametric models on

Model		Natural Questions				TriviaQA				WebQuestions			
		Total	Overlap	Comp -gen	Novel -entity	Total	Overlap	Comp -gen	Novel -entity	Total	Overlap	Comp -gen	Novel -entity
Non-parametric	RAG	44.49	75.75	30.41	37.69	56.83	87.12	47.58	47.81	45.52	80.64	33.40	31.88
	FiD	**53.13**	78.85	**40.00**	**47.74**	**67.69**	**90.39**	**58.10**	**66.23**	-	-	-	-
	DPR	41.27	71.33	25.88	33.84	57.91	82.31	46.11	58.99	42.42	73.45	31.05	31.25
	RePAQ	47.26	78.61	34.21	36.85	52.06	89.08	42.95	38.38	-	-	-	-
Parametric	T5-11B+SSM	36.59	**81.48**	17.47	12.56	-	-	-	-	44.69	81.24	35.35	25.78
	BART	26.54	76.34	5.88	3.35	26.78	78.38	11.37	10.09	27.41	70.46	13.28	8.75

Table 3: Exact Match scores for each model. "Total" refers to the overall performance on the full test set. "Overlap", "Comp-gen", and "Novel-entity" refers to the model performance on the respective subset.

NQ and WebQuestions. We manually checked the false predictions for T5-11B+SSM on the NQ overlap split, and observe that the majority are semantically correct answers, but do not exactly match the gold answer, such as ("the Outfield" v.s. "English rock band the Outfield"). The effects of the strict nature of the exact match metric has also been noted by Roberts et al. (2020) and Min et al. (2021). We note that the F1-score for these "incorrect" predictions is 44%.

Non-parametric models on Novel-entity questions For the non-parametric models, EM scores on *novel-entity* questions are relatively close to their overall total scores, with an average drop by 6.5% and 3.1% on NQ and TriviaQA respectively, with the exception of WebQuestions. The questions in WebQuestions only contain a single entity, which also tend to be high frequency entities. However, due to the very small size of the WebQuestions training set, many of these questions are considered to be in the novel-entity subset, despite containing relatively frequent entities, which, with a larger training set, would likely be classified as comp-gen questions, querying various relations regarding known entities.

Non-parametric models on comp-gen questions Surprisingly, the performance of all non-parametric models degrades significantly on the *comp-gen subset* (drop by 14.2% on NQ, 10.2% on TriviaQA and 11.7% on WebQuestions). This finding suggests that non-parametric models struggle to perform compositional generalization, whereas they handle novel question entities comparatively well. We investigate this finding in greater detail in Section 4.

Parametric models on Novel-entity and Comp-gen questions parametric model performance drops significantly on both comp-gen and novel-entity subsets, but they achieve relatively higher EM scores on comp-gen questions. This indicates that novel-entity questions are more challenging for parametric models. This makes intuitive sense, since, for entities not seen during training, parametric models will struggle to "know" enough about the entity to generate a correct answer. In such cases, we find evidence that parametric models often resort to generating answers from superficially similar training questions, with 63.2% and 53.3% of answer predictions also occurring in the training data for T5-11B+SSM on NQ for comp-gen and novel-entity questions respectively.

Implications for modeling Among the non-parametric models, FiD achieves the highest EM scores for both comp-gen and novel-entity questions. FiD aggregates multiple passages together when generating answers. In contrast, the extractive DPR reader only uses the highest-scoring passage to extract the final answer. Based on observations from a simple experiment (see Appendix A.4 for details), we believe that the NQ FiD model adopts a strategy similar to a reranker, and extracts an answer from the highest latently-relevant document.

Although without access to external knowledge but only automatically-generated QA-pairs in advance when answering questions, RePAQ still achieves higher or comparable performance as retrieve-and-read model RAG and DPR. It indicates that generating, storing and retrieving questions is a valid path in terms of model generalization.

Parametric models perform significantly worse compared to non-parametric models. BART struggles to answer any novel questions correctly, while T5-11B+SSM performs better due to much larger capacity. Petroni et al. (2019) demonstrate that language models are able to recall factual knowledge without any fine-tuning and can somewhat function as an unsupervised ODQA system. However, our experiments suggest that, large-scale language models (when fine-tuned to directly answer ques-

NQ	Total	Overlap	Comp-gen	Novel-entity
Top-20	80.1	89.5	74.7	75.4
Top-100	86.1	92.0	82.4	83.1

Table 4: Top 20 and Top 100 retrieval accuracy on NQ test set for the DPR retriever.

tions using a set of training QA pairs) struggle to answer questions about low frequency entities and relations, similar to the findings of Kassner et al. (2020) and Dufter et al. (2021).

4 How Do Non-parametric Models Generalize?

Experimental results show that the performance of non-parametric models degrades significantly on the comp-gen subsets across all datasets. In this section, we would like to thoroughly examine what the underlying challenge is for these questions. We focus on the generative QA model FiD, since it achieves the highest EM score on unseen questions, and furthermore use the NQ dataset as it has the largest annotated test set among three datasets.

Table 4 shows the top-k retrieval accuracy – which is the number of questions for which at least one passage of the top-k retrieved passages contains the gold answer. The difference in retrieval accuracy between comp-gen and novel-entity splits is relatively small ($< 1\%$), but are significantly lower than the overlap subset results. This indicates that the retriever performance is a confounding factor for the overall performance of comp-gen and novel-entity questions. Solely improving on the retriever would benefit the model greatly for the subsets requiring generalization. To isolate the behavior of the reader model, for the remainder of our analysis we only use the subset of questions for which there is at least one support passage that contains the gold answer.

4.1 Effects of Question Pattern Frequency

To measure compositional generalization on ODQA, one might ask *“How many episodes are there of Gavin and Stacey?”* and *“Who plays the doctor in Sons of Anarchy?”* as training questions and test on *“Who plays Stacey’s mum in Gavin and Stacey?”*. Although the wiki_entities of the train and test questions are different, they use the same question pattern “who plays [entity] in [entity]”. We obtain question patterns for analysis by replacing all wiki_entities in a question with the token *[entity]*, unifying the prepositions, and finally stemming each word.

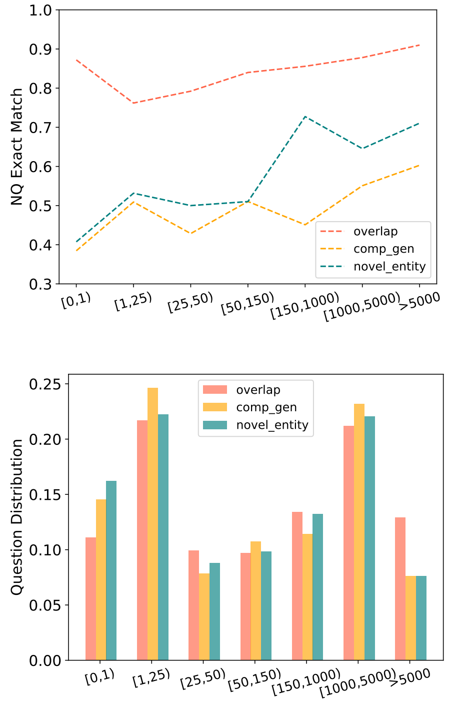

Figure 3: Plots showing the influence of question pattern frequency, where questions are binned based on their pattern frequency in the training set.

In Figure 3, questions are grouped by their pattern frequency in the training set. In the upper figure, the EM scores on all test subsets show the model is more likely to make correct predictions for more common types of questions. For less frequent patterns, the EM scores on comp-gen questions are similar to that of novel-entity questions. However, for more common question patterns novel-entity questions are more likely to be correctly answered. The lower figure shows the frequency distribution of question patterns for each subset. This plot demonstrates that the frequency distribution for each subset is similar, and thus the performance gap cannot be explained by the question pattern frequency distribution. Furthermore, this gap is non-trivial since common pattern questions take a majority of the whole test set. Under the assumption that the model could answer all patterns of questions equally, regardless of frequency, the overall performance would be improved by $\sim 11\%$. To further illustrate this observation, we sample the same number of comp-gen and novel-entity questions for each example pattern (see Figure 5). We find that the model fails more on comp-gen questions, partially because the retrieved passages do not provide sufficient information. The support passages for novel-entity questions, on the contrary, more often cover enough anchor entities. Appendix A.5 contains further details.

4.2 How do Non-parametric Models Handle Comp-gen Questions?

We would like to study if the frequency of wiki_entities in the training set affects model performance. Figure 4 plots the EM score as a function of how often a test question's wiki entity appears in a training question. We see that accuracy is *anti-correlated* with the training-set frequency with of test questions' entities. At first glance, this result seems surprising, and inconsistent with the well-known difficulty of modeling long-tail phenomena. However, the following interpretation helps to explain this apparent contradiction.

We manually inspected the questions with the most frequent wiki_entities, and find most of them are questions about countries, which is a frequent question category in the NQ training set. For example, for the question *"How many farmers are there in the USA"*, almost all the retrieved passages are highly relevant. The gold answer is "3.2 million" with the context *"There were 3.2 million farmers"*. The model, however, generates the answer "2.2 million", taken from the context *"There were 2.2 million farms…"*. Both passages come from an article titled "Agriculture in the United States", and the model is failing to draw a distinction between *farms* and *farmers*. While it is easier to retrieve relevant documents for questions with more frequent wiki_entities (Chen et al., 2021), the passages retrieved for high-frequency entities are much more likely to contain type-consistent close-negatives and distractors, making it more difficult for the model to select the correct the answer. Other questions are highly ambiguous, such as, "*What is the average salary for a US congressman*", the gold answer *$174,000* applies for the year 2012, while predicted answer *$169,300* applies for the year 2008. For NQ, the existence of high-frequency entities could be indicative of an ambiguous question. If we conduct an analysis using the NQ dev set annotations provided by Min et al. (2020), we note that 50% of questions with the entity "*US*" and 64% questions with the entity "*NBA*" are ambiguous. To quantify the impact, we note that if we match the performance of comp-gen questions with common wiki_entities to those with the unpopular wiki_entities, the accuracy could be improved with $\sim 4\%$ points.

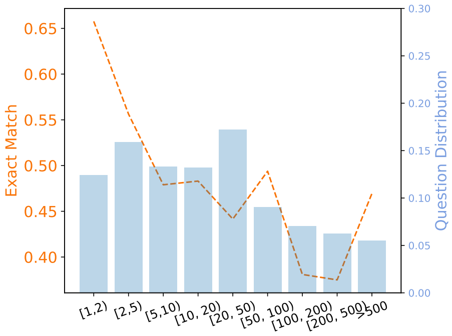

Figure 4: Plot showing the influence of the wiki_entities frequency in the question. The x-axis represents the wiki_entities frequency in the training set and we use the most frequent wiki_entities in each comp_gen question.

4.3 How do Non-parametric Models Handle Novel-entity Entities in the Question?

Although we explicitly categorize unseen questions into comp-gen and novel-entity, broadly speaking, questions with novel entities also require the model to generalize to novel compositions and thus could be considered to belong to the comp-gen category. We seek to understand if the novel entities are the main bottleneck for ODQA models, or the model can handle them well enough to process the corresponding questions appropriately. To explore this issue further, we run an ablation study, where, at inference time, we replace the novel entities in the question *and* the support passages with an entity that has been seen from the training set.

We run the inference on 100 eligible questions, and find the model rarely changes its predicted answers, despite the modification, with 73% of the predicted answers remaining unchanged. We manually verified the remaining questions and observe that some differences are due to inherent limitations of our entity-swapping process, such as errors in entity-linking (see Appendix A.6 for examples). Interestingly, we find that three altered questions give the right answers, despite originally generating incorrect ones. Given these observations, we suggest that the model learns relatively good contextual embeddings for the novel entities by exploiting the context provided by the passages. Thus, specific unseen entities are not the main bottleneck for the model to locate the desired answers.

5 Related Work

5.1 Open Domain Question Answering

Early systems relied on surface text pattern matching methods to detect answers (Ravichandran and Hovy, 2002; Soubbotin and Soubbotin, 2001). For traditional ODQA systems, linguistic experts first identify a set of question types and expected answer types using rule-based mapping methods for each type of questions (Allam and Haggag, 2012). The input question needs to be classified into a certain type or taxonomy in order to be answered (Li and Roth, 2002; Suzuki et al., 2003). This approach is sub-optimal for most realistic use-cases, as it is not possible to enumerate all possible question types.

With the introduction of deep neural networks, more recent ODQA system mostly adopt the "Retrieve-and-Read" architecture, popularized by Chen et al. (2017), retrieving relevant documents for the given question and inferring a final answer from these documents. Recent retriever models learn to encode questions and documents into dense vectors to score their similarity (Lee et al., 2019; Karpukhin et al., 2020; Khattab et al., 2020). Reader models can generally be categorized into *extractive* models that predict answer span within the document (Das et al., 2018; Lin et al., 2018; Yang et al., 2019) and *generative* that generate answers condition on the question and the retrieved passages (Lewis et al., 2020b; Izacard and Grave, 2020). Compared to traditional systems, recent ODQA models improve substantially on answer prediction (Zhu et al., 2021). However, as shown in Section 4.1, they still struggle with complicated and less common types of questions.

5.2 ODQA Model Analysis

Retrieving relevant passages for given questions is an essential component for open-book ODQA models. A broad spectrum of recent works apply transformer (Vaswani et al., 2017) models such as BERT (Devlin et al., 2019) in information retrieval (Lin et al., 2020). Following the success of the approach employing pretrained language models (Craswell et al., 2020), several work empirically study the properties of deep-learning-based retrievers. Luan et al. (2021) compare the lexical-matching abilities of these models to traditional methods such as BM25. Ma et al. (2021) and Wang et al. (2021) study their reproducibility, and demonstrate improvements by combining lexical-matching and dense retrievers. Thakur et al. (2021) introduce BEIR benchmark to study the zero-shot generalization capabilities of multiple neural retrieval approaches. Their conclusion is consistent with our findings that there are considerable room for improving generalization abilities in dense-retrieval models.

To infer answers from retrieved documents, models generally use a *reader* component implemented using a neural Machine Reading Comprehension (MRC) model. Previous work has analyzed the MRC model by crafting adversarial attacks (Jia and Liang, 2017; Mudrakarta et al., 2018), studying the difficulty of popular benchmarks (Kaushik and Lipton, 2018), and demonstrating annotation bias (Gururangan et al., 2018; Sugawara et al., 2018; Chen and Durrett, 2019). Despite the success at various datasets, there is little work analyzing the complete pipeline of question answering systems. Lewis et al. (2020c) showed that models perform substantially worse on questions that cannot be memorized from train sets. Krishna et al. (2021) found that long-form question answering (LFQA) systems do not ground their answers in the retrieved passages. In contrast, for ODQA, we observe that when we replace the retrieved passages with randomly-sampled passages at inference time, the model FiD (Izacard and Grave, 2020) largely fails to correctly answer any questions at all (see Appendix A.7 for experimental details). Gu et al. (2021) define similar generalization levels based on schemas for Knowledge Base Question Answering. However, our setting lacks a schema and our generalization categories are derived from question decomposition atoms.

6 Conclusion

We study ODQA model generalization and categorize unseen questions into three subsets: *overlap*, *comp-gen*, *novel-entity*. Treating questions as being compositional, we decompose them into atomic elements based on their semantics. It is our belief that this decomposition strategy can help future work related to question structure and unification. We evaluated several recent ODQA models on these three subsets for three popular datasets. Our experimental findings establish that novel-entity entities are not the main bottleneck for non-parametric models and we identify key factors that impact their performance. Lastly, our findings suggest specific areas to target for improvement, which in turn should lead to more robust and general ODQA models.

References

Ali Mohamed Nabil Allam and Mohamed Hassan Haggag. 2012. The question answering systems: A survey. *Science*, 2(3).

Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. 2018. Systematic generalization: What is required and can it be learned? In *International Conference on Learning Representations*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based nlp. *arXiv preprint arXiv:2106.06830*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4026–4032.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M Voorhees. 2020. Overview of the trec 2019 deep learning track. *arXiv preprint arXiv:2003.07820*.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2018. Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Philipp Dufter, Nora Kassner, and Hinrich Schütze. 2021. Static embeddings as efficient knowledge bases? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2353–2363.

Hady Elsahar and Matthias Gallé. 2019. To annotate or not? predicting performance drop under domain shift. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2163–2173.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Nora Kassner, Benno Krojer, and Hinrich Schütze. 2020. Are pretrained language models symbolic reasoners over knowledge? In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 552–564.

Divyansh Kaushik and Zachary C Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*.

Omar Khattab, Christopher Potts, and Matei Zaharia. 2020. Relevance-guided supervision for openqa with colbert. *arXiv preprint arXiv:2007.00814*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Brenden M Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *ICML*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2020c. Question and answer test-train overlap in open-domain question answering datasets. *arXiv preprint arXiv:2008.02637*.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. Paq: 65 million probably-asked questions and what you can do with them. *arXiv preprint arXiv:2102.07033*.

Belinda Z Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6433–6441.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: Bert and beyond. *arXiv preprint arXiv:2010.06467*.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. 2021. A replication study of dense passage retriever. *arXiv preprint arXiv:2104.05740*.

R Thomas McCoy, Junghyun Min, and Tal Linzen. 2020. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227.

Sewon Min, Jordan L Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, et al. 2021. Neurips 2020 efficientqa competition: Systems, analyses and lessons learned.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual meeting of the association for Computational Linguistics*, pages 41–47.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426.

Laura Ruis, Jacob Andreas, Marco Baroni, Diane Bouchacourt, and Brenden M Lake. 2020. A benchmark for systematic generalization in grounded language understanding. *Advances in Neural Information Processing Systems*, 33.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Martin M Soubbotin and Sergei M Soubbotin. 2001. Patterns of potential answer expressions as clues to the right answers. In *TREC*.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219.

Jun Suzuki, Hirotoshi Taira, Yutaka Sasaki, and Eisaku Maeda. 2003. Question classification using hdag kernel. In *Proceedings of the ACL 2003 workshop on Multilingual summarization and question answering*, pages 61–68.

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Shuai Wang, Shengyao Zhuang, and Guido Zuccon. 2021. Bert-based dense retrievers require interpolation with bm25 for effective passage retrieval.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774*.

A Appendix

A.1 Question Decomposition

Below is a random selection of question decomposition examples from the NQ dataset. In each question, $\underline{x}_{qw}$ denotes the question_word, $\underline{y}_{verb}$ denotes the verb, and the spans of other_args and wiki_ents spans are denoted by brackets. Note that these structure slots are not always fully present in the question (e.g, Q3, Q4, Q6, Q7, Q10).

As we rely on automated systems as a part of our decomposition process, this leads to the following limitations. At times, the ELQ model fails to label wiki_ents, such as for Q8 where *every light in the house* is marked as other_args. Furthermore, as seen in Q9 there is the possibility of multiple questions words being present although our approach only extracts a single question_word. Limitations such as these is one motivation for why we elected to perform manual verification for each question (Section 2.3).

1. $\underline{\text{Who}}_{qw}$ $\underline{\text{is}}_{verb}$ the [*other_args*: owner] of [*wiki_entities*: Reading Football Club]?

2. $\underline{\text{Who}}_{qw}$ $\underline{\text{died}}_{verb}$ in the [*other_args*: plane crash] [*wiki_entities*: Grey's Anatomy]?

3. [*other_args*: Cast] of [*wiki_entities*: Law & Order Special Victim Unit]?

4. $\underline{\text{When}}_{qw}$ did [*wiki_entities*: United States] $\underline{\text{enter}}_{verb}$ [*wiki_entities*: World War I]?

5. $\underline{\text{Where}}_{qw}$ are most [*wiki_entities*: nutrients] $\underline{\text{absorbed}}_{verb}$ in the [*wiki_entities*: human digestive tract]?

6. $\underline{\text{When}}_{qw}$ did the [*other_args*: government] $\underline{\text{change}}_{verb}$ the [*other_args*: retirement age]?

7. $\underline{\text{What}}_{qw}$ $\underline{\text{is}}_{verb}$ the [*other_args*: name] of the [*other_args*: gap] between [*other_args*: two front teeth]?

8. $\underline{\text{Who}}_{qw}$ $\underline{\text{sings}}_{verb}$ [*other_args*: every light in the house is on]?

9. $\underline{\text{Where}}_{qw}$ $\underline{\text{are}}_{verb}$ the [*wiki_entities*: Winter Olympics] and when do they start?

10. [*wiki_entities*: Swan Lake] [*wiki_entities*: the Sleeping Beauty] and [*wiki_entities*: the Nutcracker] $\underline{\text{are}}_{verb}$ [*other_args*: three famous ballets] by?

A.2 Question Collection for Human Verification

We use the following selection criteria to collect candidate questions for human verification. For the overlap subset, as a first step, each q is paired with each train question that shares the same answer or have answers which are a sub-sequence of q's answer. As a second step, we then require that the train question's similarity measurement score to q is over a pre-defined threshold and that they have the same wiki_entities as q. For the remaining test questions, we consider q as a candidate for comp-gen if all of its parsed elements are covered by the collection of all parsed elements in the training set. Lastly, if there exists any novel wiki_entities in q which are not present in the training set, q is considered as a novel-entity candidate.

A.3 Generalization Subset Details

As guidelines for the human annotators, we provide the following to resolve ambiguous or potentially problematic cases: 1) For overlap, we only consider questions that are superficial paraphrases and exclude those that require more complex forms of reasoning (e.g. *Who played Mark on the show The Rifleman? / Who played the boy on the show The Rifleman?*). 2) For comp-gen, all other_args in the test question must be covered in the collection of training set entities and all question_word atoms alongside with the verb must be present in the training set. However, there are questions where other_args are not covered in the training set (e.g. *Animation Resort*) or are highly specific due to the decomposition processing and thus not covered (e.g. *fourth movie* compared to *movie* or *three different types* compared to *types*) and are thus excluded from comp-gen. 3) For novel-entity, there are cases when ELQ fails to extract wiki_ents in questions because of words variation, such as *Who sang It Going to Take a Miracle?* compared to the correct wiki_ents *It's Gonna Take a Miracle*. 4) There are also intrinsic problems in the datasets, some test questions are exactly the same as train questions but paired with different answers: (*Where did Dolly Parton grow up?* with the answer *Tennessee* and *Where did Dolly Parton grew up* with the answer *Sevierville*). Following this manual verification, for Natural Questions, WebQuestions, and TriviaQA, 70.3%, 81.3%, and 69.5% of their test questions are covered in the generalization subsets respectively.

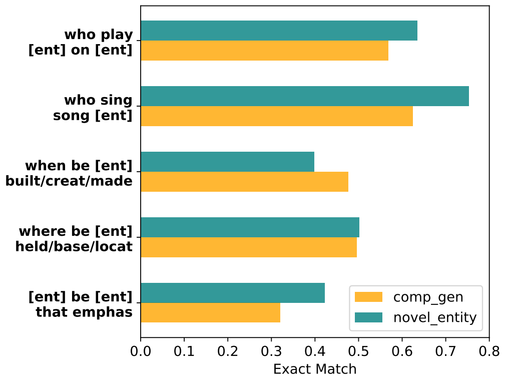

Figure 5: Examples of question patterns and EM scores for their corresponding questions. For each question pattern, we the sample same number of comp-gen and novel-entity questions. The two uppermost patterns are the most frequent (thousands of occurrences), the following two are of medium frequency (hundreds of occurrences), and the last is a novel pattern.

A.4 FiD Performance Analysis

Among the non-parametric models, FiD achieves the highest EM scores for both comp-gen and novel-entity questions. We are interested in understanding if FiD's improved performance is due to leveraging the greater amount of contextual evidence provided by multiple passages, or whether it simply generates the most frequently-mentioned plausible answer. We perform a simple experiment by first collecting 544 questions answered incorreclty by FiD, where the gold answers occur less frequently than FiD's predicted answer in the retrieved passages. We then adjust the retrieved passages so that the original predicted answer the gold answer are mentioned an equal number of times, by masking out some of the original prediction mentions. After adjusting the frequencies, we regenerate answer predictions, and observe that FiD only produces 44 correct answers out of 544. This suggests that answer mention frequency is not the governing feature for FiD when generating answers on NQ. It suggests the NQ FiD model adopts a strategy similar to a reranker, and extracts an answer from the highest latently-relevant document.

A.5 Additional Question Pattern Analyses

We sample the same number of comp-gen and novel-entity questions for each example pattern, and display the results in Figure 5. We checked several instances for the pattern "who play [ent] on [ent]", and find that the model fails more on comp-gen questions partially because the retrieved passages don't provide enough information to locate the answer. For example, for the question *"Who played Mary in Christmas with the Kranks?"* none of the retrieved passages contain both *Mary* and the movie name. The model produces the answer *Julie Gonzalo* from the passage *Julie Gonzalo Julieta [...] is an [...] actress. [She] is also known for her roles "Christmas with the Kranks"*, whereas the gold answer is *Felicity Huffman* from the passage *She also starred in [...] "Christmas with the Kranks"*. Since "Mary" is not mentioned in either passage, it is impossible to infer that the correct answer is *Felicity Huffman*. The support passages for novel-entity questions, on the contrary, more often cover both of the anchor entities (e.g. context *Little Boy Blue is an ITV drama series ... Stephen Graham was cast as Detective ...* for the question *"Who played the detective in Little Boy Blue*").

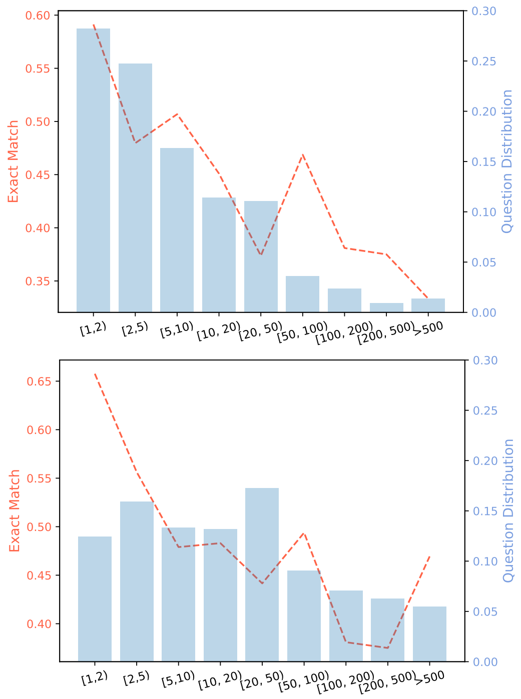

Figure 6: Influence of wiki_entities frequency in the question on model performance. *Upper*: least frequent wiki_entity frequency in the question; *Bottom*: most frequent wiki_entity frequency in the question (identical to Figure 4, but included here for the convenience of the reader).

A.6 Additional Non-parametric Generalization Analysis

When analyzing the performance impact of the frequency of wiki_entities in questions, one will have to account for the fact that there might be more than one entity present in the same question. In

Passage Processing	Total	Overlap	Comp-gen	Novel-entity
Original retrieved	53.1	78.9	40.0	47.7
50% random	53.2	78.3	39.9	48.3
99% random	55.5	74.3	46.1	54.0
100% random	3.6	5.1	2.0	3.0

Table 5: Comparison of FiD's predictions for the NQ test set, conditioned on the *originally retrieved* passages and a gradually increasing number of *randomly chosen* passages. x% means the percentage of retrieved passages are replaced with random ones. For *99% random*, the rest passage is gold passage which contains the gold answer span.

our analysis in Section 4.2 we consciously only considered the most frequent entity in a question if more than one entity was present, it is however possible that a different pattern would emerge if we would have considered the least frequent entity. To account for this, we plot both the least frequent and most frequent entity in Figure 6. We note that the same negative correlation between entity frequency and performance emerges, thus supporting the claims in our main analysis in Section 4.2.

For our experiments in Section 4.3 designed to evaluate whether a model is able to "recover" an erroneous prediction for question containing a novel entity if it is replaced with an entity observed at training time. Our experimental setup is working under the following constraints: 1) There can be only one wiki_entity mentioned in the test question, so that replacing it will not risk altering the semantics of the original question. 2) The replacement entity must not be present in the original test question or its retrieved passages. As we noted in Section 4.3, at times the novel entities in the original question may not match the corresponding mentions in the passage due to errors from the entity linking step. For instance, for the question *Who sings So Come and Dance with Me Jai Ho?* we swap the entity span "So Come and Dance with Me Jai Ho", however, this span is too wide as an entity as the correct entity would be "Jai Ho". Therefore the model is unable to match the correct song name in the passage; thus giving a different answer. Other error cases can be attributed to the granularity of the predicted answer: e.g. "624 CE" and "13 March 624 CE". We do however note that for the great majority of cases our entity-swapping procedure works as intended.

A.7 Answer Grounding in Retrieved Passages

We noted in Section 4 that we find evidence the FiD (Izacard and Grave, 2020) ODQA model does ground its answers in the retrieved passages. This observation can be contrasted to that of Krishna et al. (2021), who found that answers to long-form questions were not grounded in the passage, in that models would provide the same answer regardless of the context provided. A complete picture of the results from our experiment can be seen in Table 5. We note that when the models is fed solely random passages it fails to answer nearly all questions (3.6%). However, but provided with half gold and half random passages, it performs on par with its original performance. Lastly, we note that when presented with a single gold passage and otherwise only random passages, the model is still able to determine which passage is the gold passage and answer the question correctly – in fact, the performance even improves upon the original performance with more than more than 5% for comp-gen and novel-entity questions.

A.8 Additional Examples for three generalization subsets

Additional examples from Natural Questions are provided in Table 6, WebQuestions in Table 7, and TriviaQA datasets in Table 8.

Group	Test question	Train question
Overlap	Where does patience is a virtue come from	Where did the saying patience is a virtue come from
	Who was the killer in the movie I Know What You did Last Summer	Who was the murderer in I Know What You did Last Summer
	When was the last time Arsenal win Premier League	When was the last time Arsenal won the Premier League title
	Where does blood go when it leaves the pulmonary artery	Where does blood go after the pulmonary artery
Comp-gen	What is the most popular religion in Sweden	What is the most popular religion in Ukraine
	What are the main functions of the stem	What are the main functions of the control bus
	Who is in charge of ratifying treaties in the US	Who is in charge if president is impeached
	Cast of the Have and Have Nots play	The last episode of the Haves and Have Nots
Novel-entity	Where does wild caught *sockeye salmon* come from	When was *Sony walkman* first sold in stores
	The probability of making a *Type I Error* when retaining .. is	When was *tower of terror* built in Disneyland
	Who was the *Pinkerton Detective Agency* 's first female detective	Who played *detective Green* on Law & Order
	Where was the *world economic forum* held this year	Who holds the *world record* for 100 meters

Table 6: Example questions from NQ test set.

Group	Test question	Train question
Overlap	Which is the highest waterfall in the world	What is the tallest waterfall in the world
	In the cartoon series, what kind of dog is Scooby Doo	What breed of dog is Scooby-Doo
	Who directed the film "Gladiator", starring Russell Crowe	Who directed the film Gladiator
	Which is the largest island in Canada	What is Canada's largest island
Comp-gen	- What nationality was the painter Vincent Van Gogh	- What nationality was painter Piet Mondrian
	- What post was held by Winston Churchill during the 1926 general strike in the UK	- What role was played by Arthur Cook In the general strike of 1926
	- By population, which is the second biggest city in France	- In terms of population, which is the second largest city in Finland 1926
	- In humans, the medical condition prepatellar bursitis affects which part of the body	- The medical condition aerotitis affects which part of the human body
Novel-entity	- In *'follow that camel'*, the fourteenth carry on film, sid james was replaced by which us actor	- What was the cause of death of carmen in the opera *of that name*
	- Who has recently overtaken *brian o'driscoll* to become ireland's most capped player	- In the 2005 remake of king kong, who played the writer *jack driscoll*
	- *Shining Tor* is the highest point in which county	- *Shinto* is the main religion in which country
	- Who had a *Too Legit to Quit* tour	- Which sweets were advertised as the *Too Good to Hurry Mints*

Table 7: Example questions from TriviaQA test set.

Group	Test question	Train question
Overlap	What is the currency of Puerto Rico called	What type of currency is used in Puerto Rico
	Which countries speak German officially	What countries speak German as a first language
	What language is spoken in Haiti today	What language do Haitian speak
	What team is Hank Baskett on 2010	What team is Hank Baskett playing for in 2010
Comp-gen	What year was George W Bush elected	What is George W Bush's middle name
	What year did the Seahawks win the Superbowl	In what Super Bowl did the Seahawks face the Steelers
	Where did Queensland get its name from	From where did the Guillotine get its name
	Where was Theodore Roosevelt buried	Where is George v1 buried
Novel-entity	Where did *Andy Murray* started playing tennis	When did *Sean Murray* first appear on NCIS
	What time in *Hilo Hawaii*	Who was *Phil Harris* married to
	Where did *Bristol Palin* go to school	What team is *Chris Paul* on
	What time does *American Horror Story* air	Who made the *American Red Cross*

Table 8: Example questions from WebQ test set.