You Make me Feel like a Natural Question: Training QA Systems on Transformed Trivia Questions

Anonymous ACL submission

Abstract

Training question answering (QA) and information retrieval systems for web queries require large, expensive datasets that are difficult to annotate and time-consuming to gather. Moreover, while *natural* datasets of information-seeking questions are often prone to ambiguity or ill-formed, there are troves of freely available, carefully crafted question datasets for many languages. Thus, we automatically generate shorter, information-seeking questions, resembling web queries in the style of the Natural Questions (NQ) dataset from longer trivia data. Training a QA system on these transformed questions is a viable strategy for alternating to more expensive training setups showing the F1 score difference of less than 6% and contrasting the final systems.

1 Introduction

Question answering is a central problem in AI research. One way of understanding *why* people ask questions was explained in Rogers et al. (2023): questions come from either an information-seeking paradigm (Voorhees, 2019, henceforth information-seeking) or a probing, evaluative paradigm (Turing, 1950, probing).

While it is easy to get *questions* in the information-seeking paradigm because the asker creates questions that they do not know the *answer* to, additional annotations to find these answers are expensive. For example, Natural Questions (Kwiatkowski et al., 2019), a benchmark dataset collected by Google from questions people asked online, critically does not include the correct *answers*. Annotating answers could be more expensive than their probing counterparts, mostly written by QA writing experts (e.g., trivia members).

Moreover, while large corporations can collect large-scale *natural* information-seeking questions *at no cost*, these questions lack in quality for their ambiguity (Min et al., 2020) and false presuppositions (Yu et al., 2022). Due to these downfalls, Boyd-Graber and Börschinger (2020) argue that probing questions are more useful for building QA systems. Thus, we utilize the Quiz Bowl (QB) samples, a probing QA dataset, created by trivia experts (Section 2).[1]

This paper investigates whether and how we can transform the probing QB samples into questions that resemble natural, information-seeking questions. To this end, we propose a syntactic transformation technique NATURALIZATION that converts QB elicitations into QB-TRANS questions that resemble NQ (Section 3).

To validate the quality of QB-TRANS for training QA systems, we consider two experimental settings: zero-shot and supervised. The zero-shot setting examines whether QB-TRANS is an effective training data for a QA system when compared to NQ (Section 4). We train QA systems with QB-TRANS training data and compare the two systems on the NQ test set. Average F1 scores on NQ test set vary by less than 6%, which implies that QB-TRANS can replace NQ training data.

We also combine NQ with QB-TRANS as training data in our supervised setting (Section 5), improving F1 (tested on NQ test set) by 10% compared to training on only NQ. QB-TRANS lacks issues that plague NQ: presupposition and ambiguity (Section 7). Moreover, NATURALIZATION generalizes to other datasets. Our contributions are naturalizing of probing QB dataset into information-seeking QB-TRANS while retaining the positive traits of QB samples, thereby improving QA performance with a more affordable process. Section 9 shows how this can ensure a cheaper and more up-to-date alternative to NQ data which benefits different models and datasets.

[1] QB writers are particularly known for understanding what makes for a good QA pair; QB dataset avoids the ambiguity and false presuppositions that are often in NQ.

2 Artful but Arcane QB dataset

This section discusses why we use QB data and how different they are from NQ questions. The next section explains NATURALIZATION (Section 3).

Elicitations from QB dataset Consider this QB sample example:

> A radio mast named for this city was the world's tallest structure until the mast collapsed in 1991. This capital contains a skyscraper formerly known as the Joseph Stalin Palace of Culture and Science. A landmark called Sigismund's Column commemorates Sigismund III Vasa, who moved his capital from Kraków to this city on the Vistula River. A 1943 Jewish ghetto uprising occurred in—for 10 points—what Polish capital?

Here, clues are introduced pyramidally—harder, more obscure clues about Warsaw are sorted to appear at the first sentence (Rodriguez et al., 2021)—so that whoever knows the most about Warsaw should be able to answer the question sooner.[2]

However, we do not need this complexity. Instead, we extract the series of clues that an expert author thought was noteworthy about *Warsaw* (e.g., key sites that commemorate its history and rulers who made it the capital).

We define the source text paragraph as *elicitation*. As they are combined pieces of clues in multiple sentences, they are not grammatical or natural. Thus, we turn each clue extracted from elicitation into multiple NQ-like questions, which are short and simple. Ultimately, our goal is NATURALIZING these clues into information-seeking, *natural* questions.

Comparison with NQ datasets For each QB elicitation, we extract an average of seven clue sentences. Each sentence is 22 words on average. On the other hand, in NQ, the average sentence length is eight words (Kwiatkowski et al., 2019). The NQ questions were harvested from Google queries based on specific heuristics.[3] The number of samples from QB and NQ are comparable (QB: 112,927 elicitations and answers and NQ: 307,373 samples); however, there exists a substantial difference in cost, quality, and quantity.

For cost comparison, while the QB elicitations have answers unambiguously created by trivia authors, answers to NQ questions must be laboriously annotated by paid workers. While Google has not officially released costs, the convoluted process and the lack of reproduction since 2019 suggests that its price is high. From the QA researcher's perspective, the elicitation process is free.

For quality comparison, trivia authors who created QB elicitations understand the importance of discouraging ambiguity and false suppositions in their clues (Boyd-Graber and Börschinger, 2020) while they are prevalent in NQ. Thus, if we can faithfully elicit these clues from QB, the resulting questions may be of higher quality than NQ questions (Detail analysis is in Section 7).

Finally, for quantity comparison, because each QB elicitation contains many clues, the the size of a transformed dataset is three-fold larger than NQ. Also, while the NQ dataset may only ask a single question about a rare entity, this is not likely the case for QB: a single elicitation would produce several clues about an entity, allowing a model to understand more about each potential answer.

3 NATURALIZATION

This section outlines NATURALIZATION: converting the elicitations into multiple NQ-like questions (Figure 1).

3.1 Generating Candidates

Many of the transformations depend on an initial dependency parse (Nivre, 2010). Some parsed elicitations are statements about a target entity that do not resemble how questions are asked (e.g statements about the target entity "she was the last Queen of Hawaii" or "this element is mined from bauxite"). To transform these into questions, we find mentions coreferent with the answer.

Conjunction and Removing Clauses Given these candidates, we then extract the minimal facts that would form the basis of a question. For example, if the QB elicitation had "he wrote *Animal Farm* and 1984", this can become two facts: "he wrote *Animal Farm*" and "he wrote *1984*". Thus, we construct independent clauses by extracting spans that contain the mention ("he"), a verb ("wrote"), and one member of a conjunction (either of the two works). Similarly, we can sometimes remove clauses: "this author who graduated Eton

[2] For example, deciding it "moved his capital from Kraków to this city on the Vistula" requires the ability to decide not just what to answer, enough to answer but also *when* to answer in the quiz bowl tournament (He et al., 2016).

[3] For example, the questions start with "who", "when" or "where" followed by a finite form of "do" or a modal verb (Kwiatkowski et al., 2019)

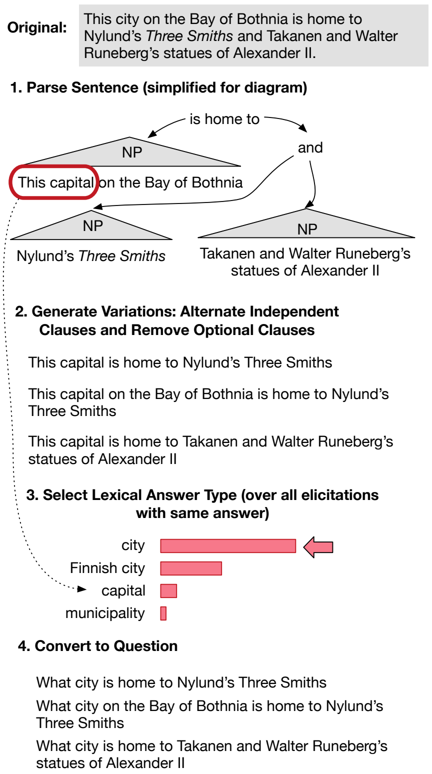

Figure 1: In the process of creating information-seeking style questions from probing elicitations, (1) we take each clue sentence from the paragraph-long QB question, and parse it. (2-3) The parsed sentences are transformed into variants, (4) that are finally turned into information-seeking questions.

College wrote *Homage to Catalonia*" can be simplified to "this author wrote *Homage to Catalonia*" (Details in Appendix, Algorithm 2).

Canonical Answer Type Next, we identify what kind of answer the question is looking for. This is important because sometimes questions written in QB's pyramidal style uses oblique references, particularly at the beginning of the question: "substance" for <u>zinc</u>, "creator" for <u>Chinua Achebe</u>, or "polity" for <u>Bangladesh</u>. However, these are rarer than the most straightforward and direct references. For example, <u>zinc</u> is most often asked about using "what element", <u>Chinua Achebe</u> with "what playwright", and <u>Bangladesh</u> with "what nation". Thus, we group all QB elicitations that have the same answer and for each answer find the most frequent string used to ask about the answer. These canonical answer types then replace the mentions in the original question.

Imperative to Interrogative The most obvious difference between QB elicitations and NQ questions is that QB elicitations are not grammatical questions: rather, they are declarative statements about the answer. For imperative statements such as "name this first prime minister of Canada", we generate a synthetic mention that makes the object of the imperative verb the question: "who was the first prime minister of Canada" by mapping the canonical answer type to its WORDNET (Fellbaum, 1998) hypernym and applying the appropriate question word (e.g., `person.n.01` maps to "who", `time_period.n.01` maps to "when"). The whole pseudocode is given in Algorithm 4 and 5.

Additional Heuristics Through observation of the linguistic and grammatical style of NQ we add additional heuristics to further improve the candidates such as **removing punctuation** and **adding subject** (full list in Appendix A).

3.2 LLM Transformation: Llama2 and GPT

As a baseline, we transform QB clues into questions by prompting GPT and LLAMA2 (Touvron et al., 2023), generative text models.[4] For a fair comparison, we follow a similar process of generating NQ-like question from QB. We extract clue sentences from the QB elicitations, then prompt GPT and LLAMA2 to produce a natural question. We first identify the lexical answer type (e.g., "this person") and ask GPT and LLAMA2 to formulate a query that could be used as a Google search. This is an example prompt for GPT and LLAMA2 (text in bold is specific to answer):

> This "sea Prince of Wales" leaves his home by hiding out in a canoe near a coral reef, and he is mistakenly called "Hedgehog" by a character who offers him a ninetieth lay, a partner of Bildad named Peleg. This is a trivia question. Turn this into shorter question of fewer than 20 words that start with "what sea", and ask about this "sea prince" in the short question. The questions should be natural as a Google query to find out what the answer to the long question is. The shorter questions you write should not include the answer, **Queequeg** and not be confusable with other answers.

.

LLAMA2 and GPT transformations do not contain all the clues or hallucinate some information (Table 1). For example, LLAMA2 did not capture the entity of interest and asked about the "sea" instead of a "character". LLAMA2 and GPT have similar generated questions (Examples in Table 8).

[4] `https://huggingface.co/meta-llama/Llama-2-7b-hf`

Original QB	QB-TRANS	LLAMA2
This "sea Prince of Wales" leaves his home by hiding out in a canoe near a coral reef, and he is mistakenly called "Hedgehog" by a character who offers him a ninetieth lay, a partner of Bildad named Peleg.	which "sea prince of wales" leaves his home by hiding out in a canoe near a coral reef, and he is mistakenly called "hedgehog" by a character who offers him a ninetieth lay, a partner of bildad named peleg.	What sea did the "sea Prince of Wales" hide out in a canoe near a coral reef?
This substance moves to the liver where it is turned into glucose in the Cori cycle.	which substance moves to the liver where it is turned into glucose in the Cori cycle.	What substance is produced when the body breaks down carbohydrates during intense exercise?
In one of this man's paintings, one character oddly uses her left hand to grasp the red-cloaked character's chin while her right hand sits at his knee	In one of which man's paintings, one character oddly uses her left hand to grasp the red-cloaked character's chin while her right hand sits at his knee?	What is the significance of the left-hand grasp and the right-hand placement in Jean Auguste Dominique Ingres' painting featuring a red-cloaked character and another woman?

Table 1: LLAMA2 converted question can introduce error in question. For example, conversion of QB clues using rule-based and large language model-based techniques (LLAMA2). In the case of LLAMA2 of the first example, the question changed the entity the QB clue is asked about. In the second example, words were removed from the question and additional random words were added, implying LLAMA2 transformation is worse than that of QB-TRANS.

4 Zero-shot QA with QB-TRANS training

We ensure we use no NQ data and evaluate on NQ test set (which disadvantages our approach).

4.1 Challenges in Zero-shot QA System

There are challenges in comparing models for zero-shot QA because some models are based on large language models (LLMs) that do not disclose training data. Thus do not know whether some zero-shot systems use NQ in their pretraining process (Shi et al., 2023a). For example, Oscar Sainz (2023); Narayanan (2023); Magar and Schwartz (2022); Sainz et al. (2023a,b) suggest that GPT-3.5 is contaminated with NQ training and development set.

One sign that these models train on NQ is that they give an abnormal probability for tokens in NQ as measured by Min K% probability (Shi et al., 2023a). The state-of-the-art LLMs have an average probability of 63% (Detail of the results in Appendix, Table 11). This indicates that these state-of-the-art LLMs has a high probability of having NQ in the training data.

Another clue that these models have used NQ for training is that they repeat NQ answers to questions even when NQ is wrong (manually detected) (Table 2); this is the clearest signal that the model has seen the NQ data's answers, as annotation errors are less likely to be by coincidence. For example, we probe GPT with time-sensitive questions that have answers no longer valid. We observe that GPT incorrectly answers those questions, with the answers included in the NQ dataset. We infer that it is likely for GPT's training data to be contaminated (Sainz et al., 2023a; Cotton et al., 2024) and can no longer be a fair candidate for the zero-shot setting experiments.

4.2 Zero-shot QA systems

Thus, we select two systems with high accuracy on traditional NQ training: Deep Passage Retrieval (Karpukhin et al., 2020b, DPR) and Retrieval-Augmented Language Modeling Framework (Shi et al., 2023b, REPLUG).These systems are trained from the ground up. **DPR** (Karpukhin et al., 2020a) extracts the answer from a context which is extracted using passage retriever models. We train DPR on the questions, answers, and context passages for the NQ-like generated QB-TRANS questions dataset (ours). In training, we generate the positive context by collecting passages that contain answer string, and negative context otherwise (Example in Appendix, Table 9). In **REPLUG** (Shi et al., 2023b), the retrieval model finds the most appropriate passage from a large corpus; then the model produces more accurate answers by augmenting retrieved information to the input context.

4.3 Training Data

We compare all of our generated datasets with the original NQ dataset (**NQ**). Our goal is to create a QA system with the same accuracy as the original NQ dataset while training on the QB-TRANS dataset, so this is an upper bound. In this zero-shot experiment, we used different percentages of QB-generated questions for training the model. We compare this traditional training regime with several training sets derived from QB-TRANS. The full results are given in Appendix, Figure 6. We compare against all transformed sentences from our syntactic-based method (**QB-TRANS**) to the LLM baseline (QB-GPT and QB-LLAMA2).

We used multiple passes when difference in dataset size. For example when the dataset size for NQ is 307k, we used multiple passes to compare against QB-TRANS dataset of size 800k.

NQ question	NQ answer (wrong)	Gold answer	GPT answer	Comment
who won the Oscar for best picture in 1976?	Rocky	One Flew Over The Cuckoo's Nest	Rocky	Rocky won the best picture in 1977 (osc, 2023).
where was held the first session of Muslim league	Dhaka, Bangladesh	Karachi	Dhaka, Bangladesh	The AIME Conference in 1906, held at Dhaka, Bangladesh, laid the foundation of the Muslim League. (mus, 2023)
Total number of death row inmates in the us	2,718	2,331	Over 2,400 people	This information is changed over periods.
Who is next in line to be the monarch of England	Charles, Prince of Wales	Prince William	Charles, Prince of Wales	The answer is outdated.

Table 2: To determine whether NQ is in the training data of GPT, we take the answers given by GPT 3.5. If the answer is the same as given in NQ dataset, we can assume it has seen those datasets.

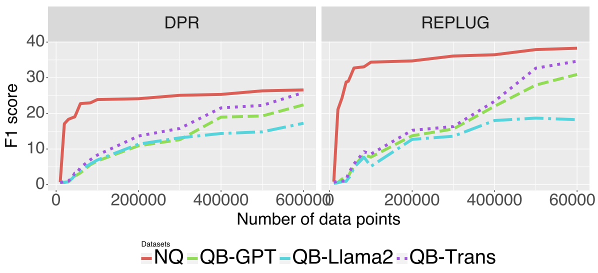

Figure 2: QB-Trans can replace NQ in training QA system and achieve accuracy close to NQ training system. **DPR**: As expected, **QB-TRANS** without any NQ data comes within 5% of a model trained on **NQ**. Training on the full QB-TRANS and evaluating it produces the highest F1 score system with DPR. This does better than transformations created by prompting a GPT and LLAMA. **REPLUG**: Again, **QB-TRANS** without any NQ data comes within 7% of a model trained on **NQ**.

4.4 Results and Analysis

Our transformations lag behind a model trained directly on NQ by only about 6% on average, while the LLM lags by over 10%. QB-TRANS data can be applied to different QA systems and achieve comparable performance (Figure 2).

LLM-based transformation (QB-GPT and QB-Llama2) performs worse than syntactic NATURALIZATION. This happens because even the worst transformed questions from the QB-TRANS dataset are better than many of the questions produced by the LLM (Table 1). Not only does the desired answer change in LLM-based transformation (it is not clear that there is a correct answer), but the answer also appears in the question (despite prompt instructions).

5 Supervised QA System with QB-NQ training data

We compare all of the naturalized datasets with the original NQ dataset (**NQ**), with the goal of having the largest NQ-like dataset.

5.1 Supervised QA systems

As the baseline, we use the top model in the NQ challenge leaderboard **ReflectionNet** (Wang et al., 2020): a MRC model for answer prediction and Reflection model for answer confidence. We also use the state-of-the-art **GENREAD** (Yu et al., 2023), which is a *generate-then-retrieve* pipeline QA system that directly generates the contextual documents by using clustering document representations. This method outperforms traditional *retrieve-then-read* methods. We also use the two retrieval-based systems DPR(Karpukhin et al., 2020b) and REPLUG (Shi et al., 2023b) from the previous section, but this time trained with QB-TRANS data along with NQ dataset.

5.2 Training Data

We train the supervised QA systems with our QB-NQ dataset, the combination of original NQ and QB-TRANS questions. We replace the QA systems' training data with QB-NQ dataset to see how our dataset performs when merged with the NQ dataset and whether our dataset can be used as an expansion of the NQ dataset. Here, QB-NQ-20, represents all of the filtered and transformed QB-TRANS dataset and 20% percent of the original NQ data. NQ examples are selected uniformly at random. We also used the same multiple passes when differences in dataset size like zero-shot setting. More detail on the formation of training questions and answers is in Appendix E.

5.3 Supervised Classifier

The generation process results in many questions that insufficiently resemble the information-seeking questions we want to emulate: some are too short or long, do not make sense, or still look too much like a probing QB elicitations. Like how Goodfellow et al. (2014) uses a classifier to filter the outputs of an automatic generative process, we identify the best examples from the above process. We use a simple logistic regression classifier (Cox, 1958) trained on the generated NQ-like examples

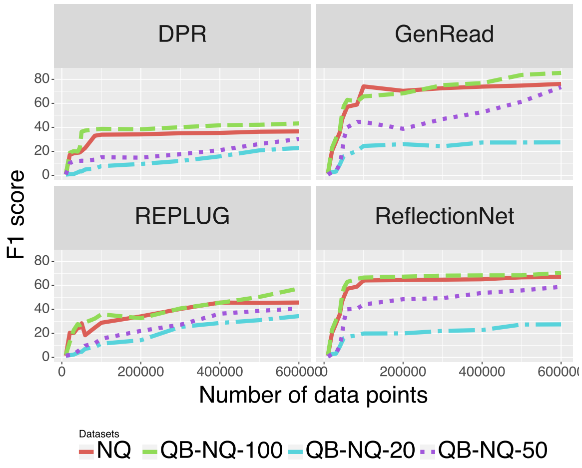

Figure 3: QB-Trans adding with NQ in training QA system can achieve F1 much higher (10% on average) to NQ training system. **DPR**: Supervised training on **QB-NQ-100** and evaluating on NQ test set produces the highest F1 score system with DPR. However, the cheaper datasets from our systematic conversion (**QB-NQ-50**), with a noisier but larger dataset, reached a substantial fraction of the F1 score. Similarly, **REPLUG, ReflectionNet and GenRead**: Again, in a supervised setting, **QB-NQ-100** data crosses the **NQ** by 10 points of a model trained on **NQ**, and adding just 50% of the NQ data (QB-NQ-50) allows the model to reach within 12% of the F1 score of the model trained on the whole **NQ** dataset.

(through the process described in the previous section) as negative examples and with real NQ examples as positive examples. To make use of the answers provided in the dataset, we designed the classifier with the answers included as a feature in the dataset.

Nonetheless, our features identify question topics and formats that occur frequently in NQ. For example, the bigram "who played", reflects NQ's emphasis on popular culture; starting questions with "how", "when", or "where" recapitulates the process for harvesting NQ; and short questions have the highest feature weight, emphasizing that NQ questions are short.

We also use early stopping with the classifier to find the optimum number of data points needed for each model. For that, we add $50k$ data at each iteration based on the classifier and test it on NQ dev set until the F1 score continues to increase. When the score starts to drop we continue it for five more iterations to avoid local minima. If F1 again starts to increase, we continue. Otherwise, the data number that has the best F1 score on the dev set is chosen as the optimal train set.

Models	Datasets			
	NQ	QB-NQ-100		
		No classifier	With classifier	
			no early stopping	*early stopping*
DPR	39.23	43.54	46.21	**49.12**
REPLUG	45.75	55.29	49.12	**57.56**
ReflectionNet	64.01	68.36	73.89	**75.87**
GenRead	74.31	79.56	**80.03**	78.01

Table 3: The best F1-score is reported here. The classifier with early stopping helps us to find out the optimal number of data points needed for the model.

5.4 Result and Analysis

We argued that using transformed QB-TRANS data would be cheaper than using NQ data (which is expensive) to gather answers. What if we have access to a *fraction* of the NQ data? Finally, given the best configuration of the previous experiment, we add a small amounts of NQ data to see how much is needed to recreate the best NQ result. Adding half of the NQ brings parity to the result. Therefore, our experiments show the effectiveness of QB-TRANS dataset as an alternative of NQ dataset in the zero-shot setting and an expansion of NQ dataset in supervised QA systems. Similar results can be seen in all the systems (Figure 3). ReflectionNet and GenRead have higher F1 score than DPR and REPLUG because of their usage of large language models and ensemble models in training. No data in the training process is changed. The result is summarised in Table 3.

6 Answer Equivalence in Zero-shot and Supervised Training

Thus far, we focused on ensuring that the transformed questions resemble the target NQ data as much as possible but did not consider the answers. To fully emulate NQ data, the answers need to be comparable. Thus, we expand the answer set provided in the QB dataset (which typically is more formal and verbose than NQ) with the WikiData answer equivalence sets from Si et al. (2021) for both training and evaluation.

For example, NQ has a question "Where do the greasers live in the outsiders?" with the correct answer set comprised of {"Tulsa", "Oklahoma"}. However, if the QA system answers "tulsa", "Oklahoma", it will be considered as incorrect in the exact match. Thus, we apply an answer equivalence system to change the answer set to {"Tulsa", "Oklahoma", "ttown", "Tulsa", "tulsa oklahoma",

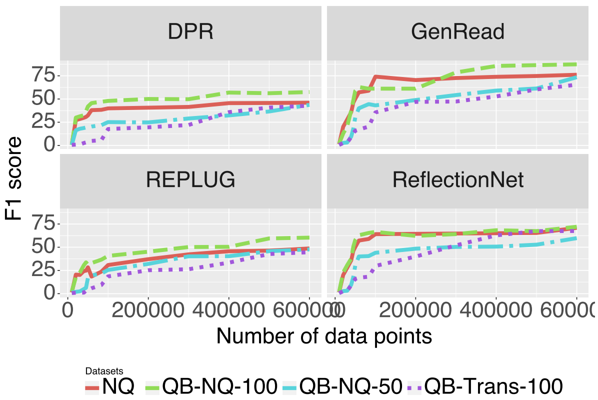

Figure 4: **With answer equivalence**: Again, **QB-NQ-100** data crosses by 12% on average of a model trained on **NQ**, and adding just 50% of the NQ data allows the model to reach within 7% of the whole NQ with answer equivalence. **QB-TRANS-100** comes within 5% points of model trained on NQ.

"wagoner county Tulsa city"}. After adding answer equivalence in the supervised setting, the F1 score for QB-NQ increased by 12% from NQ which is 3% more than systems without answer equivalence. Moreover, the F1 score for QB-NQ-50 is much closer (2% improvement) to NQ than they were without answer equivalence. In zero-shot setting, the F1 score for QB-TRANS is 5% less than the F1 score for NQ (without answer equivalence F1 score was 6% less than NQ) (consistent with results in Si et al. (2021)) (Figure 4).

7 Analysis of Transformed Questions

7.1 Quality of Generated Data

To analyze the quality of our dataset, we use CREPE (Yu et al., 2022) to identify false presuppositions (Table 4). The percentage of presuppositions present in our dataset is less than NQ.

NQ has has more ambiguous questions detected using Min et al. (2020)'s AmbigQA binary classifier and GPT-3.5 (Table 4). An example of an ambiguous question from NQ, *"How many nominations does Game of Thrones have?"* This question can ask about the number of nominations "Game of Thrones" has across all its seasons, or it can ask about any particular season or award ceremony. Therefore, no precise answer can be given without additional context. On the other hand, QB elicitation ensures each clue points to a unique object without any ambiguity.

Dataset	Size	% of Presupposition	% of Ambiguity using GPT-3.5	% of Ambiguity using AmbigQA
NQ	307373	21	63	68
QB-Trans	800000	27	27	25

Table 4: The percentage of harmful presupposition and ambiguous questions in NQ and QBTrans dataset. QB-Trans has fewer presuppositions and significantly fewer ambiguities than NQ.

7.2 Transformation Error Analysis

Not all of the original elicitations are transformed correctly. Consider this original clue from elicitation:

> This author created a character who smokes a cigarette before the body of his dead mother, and who vacations with his friend Raymond and shoots an Arab on the beach.

The heuristic "split conjunction" and "no wh-word" are applied and generate questions "This author created a character who smokes a cigarette before the body of his dead mother,", "what author vacations with his friend Raymond" and "what author shoots an Arab on the beach". The 2nd and 3rd questions are incorrect. This happens because there is an error in finding relative clauses when splitting via conjunction. In the future, we will detect these sorts of questions earlier where the transform technique will not be directly applicable via the dependency parse tree.

7.3 Cost of Heuristics and Generalization

Our process took several iterations to refine the heuristics. It took less than a hundred hours. However, all these heuristics can be directly applied to other pyramidal and clue-based question-answering datasets and generate NQ-like data at a cheaper cost without going through each clue manually.

To show the generalization of our heuristics, we apply the heuristics to different datasets. For example, *Jeopardy!* has an elicitation:

> This small, red summer fruit develops tiny seeds on the outside and often tops shortcake.

After applying the heuristics described in Section 3.1 the question becomes

> Which small, red summer fruit develops tiny seeds on the outside and often tops shortcake?

We apply these heuristics to similar clue-based datasets *Jeopardy!* (Jeo, 2024), *TriviaQA* (Joshi et al., 2017a), *HotpotQA* (Yang et al., 2018) and Japanese dataset *AI King* (AIk, 2024). Examples of the original questions from these datasets and transformed questions after applying our heuristics are in Appendix Table 12 and 13. Figure 5 shows

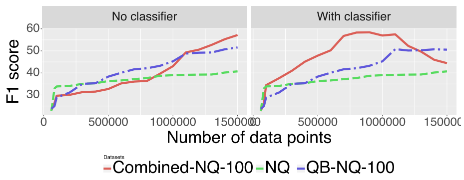

Figure 5: **No classifier:** The combined dataset shows similar performance initially with the model trained on NQ and QB-NQ. However, when we increase the data point, it goes 12% higher than the model trained only on NQ. With the **classifier**, the classifier chose the training data to resemble NQ. Therefore, the data selected earlier produces a better F1 score. However, after 110k data points, the performance starts to deteriorate. That means the data we add does not resemble NQ after that.

Models	NQ	Datasets: QB-NQ-100-Jeopardy-TriviaQA-AI King-HotpotQA		
		No classifier	With classifier	
			no early stopping	*early stopping*
DPR	39.23	52.20	**57.48**	53.54
REPLUG	45.75	58.35	57.10	**60.92**
ReflectionNet	64.01	75.91	77.96	**79.89**
GenRead	74.31	80.98	82.90	**85.87**

Table 5: The best F1-score on NQ test is reported here. The classifier with early stopping based on NQ dev helps us to find out the optimal number of data points.

the application of heuristics to other datasets can generate larger datasets and this combined dataset (COMBINED-NQ-100) can improve the F1 score for DPR. We can significantly increase the size of datasets by applying these heuristics automatically to different language and domain datasets which can increase the system's F1 score compared to the system solely trained on NQ. The results of these datasets are in Table 5. Table 10 shows the percentage of error our heuristics have while applying to different domain and language datasets is less than 1%. Our heuristics can also detect errors (e.g. ill-formed sentences, ambiguous clues about the entity, etc.) in the datasets. For example, in the *Jeopardy!* elicitation "Hits hard", it is not possible to answer that without more context. Our heuristics can be applied to identify them.

8 Related Work

8.1 Generating Questions

Given the expense of gathering these data, an obvious alternative is to generate your data. While we transform one question format into another, Probably Asked Questions (Lewis et al., 2021, PAQ) transforms source documents into questions that *could* be asked. These questions are more formulaic than the questions carefully crafted by trivia experts in the QB dataset, but an obvious extension would be to see if PAQ questions could help augment the results here. Another class of transformed questions are translated questions that convert datasets like SQUAD into multiple languages (Carrino et al., 2020; d'Hoffschmidt et al., 2020). A frequent research thrust has been to create methods to generalize these datasets, either by merging datasets together (Artetxe et al., 2019; Khashabi et al., 2020) or by QA-driven slot-filling (Du et al., 2021b) or event extraction via QA (Lyu et al., 2021) by creating algorithms that explicitly generalize (Munteanu et al., 2004; Munteanu and Marcu, 2005). More related work is in Appendix, Section C.

8.2 Transforming Questions

Our approach of transforming the form of QB elicitations is inspired by a long line of research. Machine translation models are used to transform questions to resemble the text where the answer would be found (Wang et al., 2007) or to transform a context-dependent question into a question that more closely resembles NQ (Demszky et al., 2018).

9 Conclusion and Future Work

Transformed NQ-like questions from the QB data is an alternative to expensive datasets like NQ. The transformed data itself is not as good as NQ by itself, but is competitive; this is a reasonable option if the resources are not available to curate a dataset like NQ.NQ is used text summarization, document retrieval, alignment along with benchmark of QA evaluation. However, the dataset is getting old with absolute questions and out-of-date answers. If there is a budget to create a dataset comparable to NQ, a small amount of this data augmented with transformed data from a dataset like QB can surpass a model trained on the NQ dataset alone. This can act as a continuous flow of new natural questions. Moreover, there are some methods like reinforcement learning from human feedback (RHLF) that uses NQ along with other datasets (Li et al., 2023; Feng et al., 2023) or create new datasets aligning NQ with other datasets for LLMs (Yang, 2023). Our work shows that there are additional sources of information that are cheaper and more recent that can feed into these datasets instead of NQ. For future work, we can apply this conversion technique to other languages' probing dataset (Han et al., 2023) where transformation heuristics can be learned using human data.

10 Limitations

Focus on Natural Questions We focus on NQ, a popular and respected dataset. It contains real user questions from Google on a variety of topics and they are natural queries. This diversity helps in training QA models and is suitable as a benchmark for the evaluation of QA systems. Other datasets are different, and we do not know how well our transformations would generalize to other datasets. However, we suspect that similar transformations would also succeed.

Errors hidden by Correct Answers While our transformed data often gets to the right answer, we have not systematically verified that the produced questions are themselves correct. It could be that enough of the necessary contents within the conversions remain that systems can reach the correct answer but that the questions contain errors (either factual or grammatical). From our inspection of the questions, we do not believe this to be the case, but a systematic evaluation would be needed to confirm this. However, this would dramatically raise the cost of the dataset, obviating one of the motivations for this approach.

Distribution Shift QB and NQ have very different distributions: QB is more academic, while NQ has more questions about sports and pop culture. Thus, solely evaluating on NQ potentially says little about how well our conversion process works for the topics that are over-represented in QB compared to NQ. While NQ does have some questions about literature and science, they are under-represented; it could be that our transformations are particularly brittle on questions about equations or works of fiction but NQ evaluation does not expose that weakness.

Ethical Considerations

The most important ethical consideration of this paper is that we are using the data from the trivia community to train a model. In contrast to datasets like SearchQA (Dunn et al., 2017) or TriviaQA (Joshi et al., 2017b) where it is unclear how the original trivia authors feel about the use of the data, the QB community explicitly welcomes the sharing and dissemination of the data to train QB players: datasets are covered by a creative commons license (and the norm of sharing indeed predates the formal creation of creative commons). While computer QA systems are a different kind of trivia player (machine rather than human), we believe that this would be in the spirit of the community.

References

2023. All-India Muslim League. https://en.wikipedia.org/wiki/All-India_Muslim_League.

2023. Experience over Nine Decades of the Oscars from 1927 to 2024. https://www.oscars.org/oscars/ceremonies/1976.

2024. AI King Japan Quiz AI Championship. https://sites.google.com/view/project-aio/home?authuser=0.

2024. Jeopardy! dataset. https://huggingface.co/datasets/jeopardy-datasets/jeopardy.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Pratyay Banerjee and Chitta Baral. 2020. Self-supervised knowledge triplet learning for zero-shot question answering. *arXiv preprint arXiv:2005.00316*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Association for Computational Linguistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. 2024. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, 61(2):228–239.

David R Cox. 1958. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Martin d'Hoffschmidt, Wacim Belblidia, Tom Brendlé, Quentin Heinrich, and Maxime Vidal. 2020. Fquad: French question answering dataset. *arXiv preprint arXiv:2002.06071*.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathy Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2021a. Glam: Efficient scaling of language models with mixture-of-experts. *CoRR*, abs/2112.06905.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Xinya Du, Luheng He, Qi Li, Dian Yu, Panupong Pasupat, and Yuan Zhang. 2021b. Qa-driven zero-shot slot filling with weak supervision pretraining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 654–664.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Falcon. 2024. Falcon 7B. https://falconllm.tii.ae/falcon-models.html. [Online; accessed 8-May-2024].

C. Fellbaum. 1998. *WordNet : An Electronic Lexical Database*, chapter A semantic network of English verbs. MIT Press, Cambridge, MA.

Tao Feng, Zifeng Wang, and Jimeng Sun. 2023. Citing: Large language models create curriculum for instruction tuning. *arXiv preprint arXiv:2310.02527*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

HyoJung Han, Jordan Boyd-Graber, and Marine Carpuat. 2023. Bridging background knowledge gaps in translation with automatic explicitation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9718–9735, Singapore. Association for Computational Linguistics.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In *International conference on machine learning*, pages 1804–1813. PMLR.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017a. 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, Vancouver, Canada. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017b. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. 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, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020b. Dense passage retrieval for open-domain question answering.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

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. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Lei Li, Yekun Chai, Shuohuan Wang, Yu Sun, Hao Tian, Ningyu Zhang, and Hua Wu. 2023. Tool-augmented reward modeling. *arXiv preprint arXiv:2310.01045*.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. *arXiv preprint arXiv:2203.08242*.

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, Online. Association for Computational Linguistics.

Dragos Stefan Munteanu, Alexander Fraser, and Daniel Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 265–272.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Arvind Narayanan. 2023. Gpt-4 and professional benchmarks: the wrong answer to the wrong question. `https://www.aisnakeoil.com/p/gpt-4-and-professional-benchmarks`.

Joakim Nivre. 2010. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152.

OpenOrca. 2024. OpenOrca - Mistral - 7B - 8k. `https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca`. [Online; accessed 8-May-2024].

Iker García-Ferrero Julen Etxaniz Eneko Agirre Oscar Sainz, Jon Ander Campos. 2023. Did ChatGPT cheat on your test? `https://hitz-zentroa.github.io/lm-contamination/blog/`.

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

Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. 2020. Blockwise self-attention for long document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2555–2565, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pedro Rodriguez, Shi Feng, Mohit Iyyer, He He, and Jordan Boyd-Graber. 2021. Quizbowl: The case for incremental question answering.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Computing Surveys*, 55(10):1–45.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023a. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023b. Did chatgpt cheat on your test?

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023a. Detecting pretraining data from large language models.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023b. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Chenglei Si, Chen Zhao, and Jordan Boyd-Graber. 2021. What's in a name? answer equivalence for open-domain question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9623–9629, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hao Sun, Xiao Liu, Yeyun Gong, Anlei Dong, Jingwen Lu, Yan Zhang, Daxin Jiang, Linjun Yang, Rangan Majumder, and Nan Duan. 2023. Beamsearchqa: Large language models are strong zero-shot qa solver.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the limits of ai through gamification. In *Proceedings of Advances in Neural Information Processing Systems*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

A. M. Turing. 1950. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460.

Ellen M. Voorhees. 2019. pages 45–69. Springer International Publishing, Cham. [link].

Mengqiu Wang, Noah A Smith, and Teruko Mitamura. 2007. What is the jeopardy model? a quasi-synchronous grammar for qa. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 22–32.

Xuguang Wang, Linjun Shou, Ming Gong, Nan Duan, and Daxin Jiang. 2020. No answer is better than wrong answer: A reflection model for document level machine reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4141–4150, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, and Pawan Sasanka Ammanamanchi. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Hui Yang, Lekha Chaisorn, Yunlong Zhao, Shi-Yong Neo, and Tat-Seng Chua. 2003. Videoqa: question answering on news video. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 632–641.

Jianxin Yang. 2023. Longqlora: Efficient and effective method to extend context length of large language models.

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

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2023. Generate rather than retrieve: Large language models are strong context generators.

Xinyan Velocity Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Crepe: Open-domain question answering with false presuppositions. *arXiv preprint arXiv:2211.17257*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big bird: Transformers for longer sequences.

Hang Zhang, Yeyun Gong, Yelong Shen, Weisheng Li, Jiancheng Lv, Nan Duan, and Weizhu Chen. 2021. Poolingformer: Long document modeling with pooling attention. In *International Conference on Machine Learning*, pages 12437–12446. PMLR.

A Heuristics List

Through observation of the linguistic and grammatical style of NQ we add additional heuristics to further improve the candidates such as **removing punctuation** and **adding subject**:

- **Removing punctation**: Natural questions typically do not include punctuation, so we remove punctuation at the boundary of a generated question.
- **Adding subject**: If a question is missing a subject (e.g., "wrote *Burmese Days*", we add "which" `answer_type` (in this example, author) to the beginning of the question.

Full list of heuristics in Table 6 and 7.

B Zero-shot QA with QB-TRANS Data

B.1 What is a zero-shot system?

Zero-shot systems enables the models to answer the questions without explicitly trained on them. Under zero-shot setting for the NQ dataset, there can be no training on NQ data– not with questions and their answers and not with their contexual documents. Therefore, when given any NQ test data, the zero-shot systems directly encode the given question and predict the answer. A question q is given to the model as the input. Based on that input, the model generates the answer a denoted by $p(a|p, \theta)$ where θ is the model parameters (Yu et al., 2023).

The state-of-the-art zero-shot QA system ALLIES (Sun et al., 2023) framework generates additional questions through an iterative process. In this process an LLM is used to generate queries based on existing query-evidence pair and score the answer. This iteration process continues until the score reaches a predefined threshold. Therefore, this system decomposes the original question into multiple sub-questions and achieves state of the art performance on zero-shot setting for NQ dataset. Another state-of-the-art zero-shot model GENREAD Yu et al. (2023) uses large language model InstructGPT (Ouyang et al., 2022) to directly generate contextual documents from a given question.

B.2 Min K% probability

To design a fair zero-shot system to compare NQ with QB, we first detect whether NQ data exists in the training data of an LLM by using Shi et al. (2023a)'s Min K% probability technique. This technique utilizes minimum token probabilities of a text for detecting data in pertaining. The hypothesis is that a member example in training data does not have words with a high negative log-likelihood. The average log-likelihood of K-% tokens is computed using

$$Min\text{-}K(\%)Prob(x) = \frac{1}{E} \sum_{x_i \in Min\text{-}K\%(x)} logP(x_i|x_1, \ldots x_{i-1}) \quad (1)$$

After feeding in an NQ sample into the model, we use the technique to yield Min K% probability by taking k% tokens with minimum probabilities with K=60 and calculating their average log-likelihood. Based on the hypothesis in Shi et al. (2023a), if the log-likelihood is high, then NQ is likely to exist in the model's training data.

B.3 DPR Training

The passages that contain any of the answer strings are positive examples, while the passages that do not are negative examples. One example is shown in Table 9.

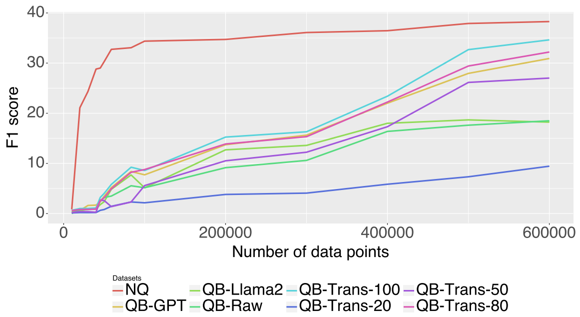

Figure 6: QB-Trans can replace NQ in training QA system and achieve accuracy close to NQ training system. As expected, **QB-Trans-100** without any NQ data comes within 5 points of a model trained on **NQ**. Training on the full QB-Trans and evaluating it produces the highest accuracy system with DPR. However, the percentage of that dataset from our systematic conversion (**QB-Trans-80**) reaches a substantial fraction of the accuracy. This does better than conversions created by prompting a LLM.

B.4 Zero-shot Training and Results

We use individual elicitation sentences from the QB dataset *without* any transformation: **QB-Raw**. While we expect this to do poorly, it shows how much our transformation improves upon the original dataset.

C Related Work

C.1 An Explosion of Datasets

The last few years have seen a flurry of datasets. Some of these datasets are created at great expense through crowdsourcing to capture common sense, numerical reasoning, visual QA (Antol et al., 2015), video QA (Yang et al., 2003), common sense questions (Talmor et al., 2021) or multicultural questions (Clark et al., 2020); Rogers et al. (2023) gives a thorough summary. Less common are datasets focusing on found data, although there is nonetheless a panoply of questions harvested from educational resources, civil service exams, users, and trivia games.

C.2 Large Language Models and Transformer-based Models

Due to the increasing sequence length, transformer uses sparse attention to handle the complexity of long document modeling (Zhang et al., 2021). In this method, each token is made to attend more important context or local context (Qiu et al., 2020). Another approach uses sliding window pattern to capture local information that includes Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2021). Lastly, PoolingFormer (Zhang et al., 2021) uses full self-attention into two-level attention schema–first one works as a sliding window attention pattern and the second level increases the receptive field. Wang et al. (2020) uses machine reading comprehension (MRC) model for answer prediction and a Reflection model for answer confidence. This achieves state-of-the-art performance on the NQ dataset in the leaderboard of NQ challenge.

C.3 Zero-shot QA

In a zero-shot setting, the large language model is used to generate new questions. In BeamSearchQA (Sun et al., 2023), new questions are generated using LLM by iterative refining and expanding the scope of the question to achieve a state-of-the-art EM score of 38.0, there are some approaches without the retriever. The in-context learning approach is applied using GPT-3 (Brown et al., 2020), cost-efficient Generalist Language Model (GLaM) GPT-3 (Du et al., 2022), instruction-tuned model (Wei et al., 2021) in zero-shot setting. Self-supervised knowledge learning is applied in zero-shot QA, for example, heuristic-based graph (Banerjee and Baral, 2020). However, in our work, we are creating nq-like questions from qb questions. The main difference between our work from the previous work is that we are using a different dataset to train the model in a zero-shot to make it compatible with the NQ dataset. With a proper classifier and carefully chosen heuristics, we introduce a conversion of different domain datasets as a replacement of the NQ dataset.

D Comparison of LLMs and Error in Transformation

D.1 GPT vs Llama2

We use llama baseline because of the cost efficiency. Both GPT and Llama2 showed similar conversion(Table 8). However, Llama baseline results are comparable to the GPT models. For example, training with the first 10000 examples ends with an accuracy of 0.58 for GPT and 0.45 accuracy for Llama2. Similarly, when we have 50000 samples for both models, the accuracy is 3.13 for GPT and 2.64 for Llama2. We can see both the language models perform worse than the rule-based conversion in the QA systems. That is why we can say, the rule-based system (**QB-Trans**) performs better irrespective of language model choice as the baseline (Figure 6).

E Answer Formation in QB

We also transform answers from the QB dataset to look like the NQ data. For example, one of the QB questions after transformation "Which ethnic group's language and customs were adopted by a majority of the uru people?" with the answer "Aymara people (the Quechua were the larger group targeted by the genocide)". However, if we observe the NQ answer list, there is no description given using the parenthesis. Therefore, we convert the answer set to also include "Aymara people" to make the answer set look like NQ formatted.

F Process of Application of heuristics

We have applied all the heuristics to all the questions with some precondition to determine the applicability of those heuristics. For example, when we apply "remove conjunctions" heuristics, we determine whether that particular question has a conjunction (via a dependency parse). If it has a conjunction, only then that heuristics will be applied. Otherwise, the question goes to the next heuristics unchanged. Similarly, for "Imperative to Interrogative" heuristic checks whether the subject of that question is imperative and if it is, converts it to interrogative.

Algorithm 1 Transform QB Questions to NQ-like Question

```
Split each clue in QB questions into QB elicitation (QB_E) by splitting them through period(.)
procedure APPLY HEURISTICS FOR TRANSFORMER(QB_E)
    Heuristics list (H)={Split Conjunction, Imperative to Integrative, No Wh-words, . . . }
    for each QB_e ∈ QB_E do
        for each heuristics ∈ H do
            AppliedHeuristic = PreCondition(QB_e)    ▷ Apply PreCondition to see whether that heuristic can be applied to QB_e
            if AppliedHeuristic is True then
                QB_e = heuristics(QB_e)
                QB_e = PostCondition(QB_e)    ▷ Apply PostCondition to check for syntax errors in the heuristics application
            else
                QB_e is unchanged
            end if
        end for
    end for
end procedure
```

Algorithm 2 In transforming QB clues into NQ-like questions, we split the clues via conjunction and construct two independent clauses by splitting them.

```
procedure POS(word)
    Return parts of speech of word
end procedure
procedure DEP(word)
    Return dependency of word in parse tree
end procedure
procedure POSITION(word)
    Return position of word in parse tree
end procedure
Flag = Check if question has conjunctions
if Flag is True then
    Parse(q) = parse tree for the question
    root verb = [x ∈ Parse(q) if PoS(x) is "VERB" and there is no ancestors for x in Parse(q)]
    verbs = [x ∈ Parse(x) if PoS(x) is "VERB" and x.head ∈ root verb]
    for verb ∈ verbs do
        for child ∈ verb.children do
            if Dep(child) is 'cc' and PoS(child) is coordinating conjunction then
                verb conj.add((verb, child))
            end if
        end for
    end for
    for verb, conj ∈ verb conj do    ▷ Check to see if this is the second verb and if it has no ancestors
        if Position(verb) > Position(verbs[0]) and if there are no ancestors for the verb in the Parse(q) then ▷ If so, we have two independent clauses, so yield the two parts on either side of the conjunction
            First question= x.text for x in parse if Position(x) < Position(conj))
            Second question = x.text for x in parse if Position(x) > Position(conj))
        else if Position(verb) < Position(verbs[-1]) and Dep(verbs[-1]) is "conj" then    ▷ Otherwise, if this verb is child of another verb with "conj" relation, we can have two sentences with the same subject, so get what came before verb and does not modify verb
            left tokens = [x for x in parse if Position(x) < Position(verb) and not (x.head == verb and (PoS(x) is "ADVERB" or "AUX"))]
                                                                        ▷ Get possible completions
            first verb = [x for x in parse if x.position < conj.position and not x ∈ left tokens]
            second verb = [x for x in parse if x.position > conj.position]
            First question =x for x in left tokens + first verb)
            second question = x for x in left tokens + second verb
        end if
    end for
end if
```

Heuristic	Purpose	Example before Heuristic	Example after Heuristic
substitute non answer pronouns	Substitute non answer pronouns to noun+possesion.	she founded Carthage and reigned as its queen from 814-759 BC	she founded Carthage and reigned as carthage's queen from 814-759 BC
clean marker	Remove punctuation patterns at the beginning and the end of the question.	which german philosopher is this philosopher wrote a work , . "	which german philosopher also wrote glowing reviews of which german philosopher's own works in ecce homo
drop after semicolon	Remove contents after semicolon in NQlike.	which molecule is this compound 's presence can be quantified in spectrophotometry by observing an intense absorption peak at 255 nanometers ; that peak is the	which molecule 's presence can be quantified in spectrophotometry by observing an intense absorption peak at 255 nanometers
convert continuous to present	Change the first verb to normal tense if it is in continuous tense.	which particle consisting of a charm quark and an anti - charm quark	which particle consists of a charm quark and an anti - charm quark
fix no wh words	Convert "this" to "which"+answer_type when there's no "wh-" words.	this play begins with the protagonist arriving at the elysian fields to see her sister stella	which play begins with the protagonist arriving at the elysian fields to see her sister stella
replace this is	Replace "this" to "which"+answer_type within "this is" pattern.	this is the first party name , followed by kraemer , in that supreme court case , which held that racially restrictive covenants are unconstitutional	which name the first party name , followed by kraemer , in that supreme court case , which held that racially restrictive covenants are unconstitutional
replace which with that	Convert "which" to "that" and check if no "which" present anymore, if so, convert "this" to "which".	michael green is a current professor at this university , which is where watson and crick discovered dna 's structure	michael green a current professor at which university , that is where watson and crick discovered dna 's structure
add question word	Adding "which"+answer_type when no "wh-" words present.	a chamberlain named cleander was killed on the orders of marcia , a mistress of this man who was involved in the plot that eventually assassinated him and replaced him with pertinax	a chamberlain named cleander killed on the orders of marcia , a mistress of which man who was involved in the plot that eventually assassinated him and replaced him with pertinax
add subject	Add "which"+answer_type at the beginning when question starting with VERB/AUX and missing the subject.	were refused real employment because of " logical discrimination , " an excuse which belied the employers ' fear of their " death taint	which se people were refused real employment because of " logical discrimination , " an excuse which belied the employers ' fear of their " death taint
fix what is which	Remove "what is" from "what is which".	what is which desert lying mostly in northern china and mongolia	which desert lying mostly in northern china and mongolia
remove end BE verbs	Remove "is/are" at the end of NQklike questions.	which jewish holiday is that hymn is	which jewish holiday is that hymn
remove extra AUX	Remove extra auxiliary words.	which number is it is the base for solutions to the differential equation	which number is the base for solutions to the differential equation
remove patterns	Remove bad patterns in NQlike.	which irish playwright is andrew (*) undershaft	which irish playwright is andrew undershaft
remove rep subject	remove repetition of the subject "is this".	which goddess is this goddess is considered a daughter of ra	which goddess is considered a daughter of ra
remove BE determiner	Change is his/is her/is its to 's.	which greek goddess's is her wedding night lasted three hundred years	which greek goddess's wedding night lasted three hundred years
remove repeated pronoun	Removes repeated pronouns like "which character who is", "is who is".	which character who is the character who never appears to linus in a peanuts halloween special	which character never appears to linus in a peanuts halloween special

Table 6: List of Heuristics

Heuristic	Purpose	Example before Heuristic	Example after Heuristic
fix no verb	Ensure there's at least one verb per question.	which greek god wielding chief greek god	which greek god is wielding chief greek god
add space before punctuation	Add space before punctuation because in NQ there's space before all types of punctuation	which greek goddess's wedding night lasted three hundred years	which greek goddess 's wedding night lasted three hundred years
rejoin whose	replace "who's" with "whose"	which wife who 's kidnapping by paris began the trojan war	which wife whose kidnapping by paris began the trojan war

Table 7: List of Heuristics.

Algorithm 3 No Wh-words: In converting question with for No Wh-words we need to introduce wh-words

```
Flag = Check if question has no wh-words
if Flag is True then                                        ▷ If no wh-words found in the question
    answer type=Find the canonical type of the answer for the question
    if question contains "this" then
        final question= replace "this" with "which" in the question
    else if If the subject of the question is pronoun then
        final question= replace the subject of the question with "which" + answer type in the question
    else
        final question=add "which" + answer type at the beginning of the question
    end if
end if
```

Algorithm 4 Heuristics for Imperative to Interrogative: If the question starts with verbs like "name," "give," or "identify", it converts it to standardized imperative question form.

```
Imperative Pattern = {(ftp | FTP | Ftp) (give | identify | name) (this | these) }, {(For | for) (ten | 10 | 20 | 5 | 15) (Points | points
    | points) (give | identify | name) (this | these)}
Flag = Check if the clue has the imperative pattern
if Flag is True then
    answer type=cannonical answer type for the question
    verb position = find the minimum position of verbs ["name", "give", "identify"] in the parse tree
    head = the head of the verb using verb position in the parse tree          ▷ Get the first noun after the verb
    if There is a relative clause in the children for the head in the dependency for the parse tree then
        relative head = relative clause's head from the parse tree              ▷ Find the relative clause head
        relative head = first element in relative head list
        continuation = concatenate text from parse starting at relative head's left edge + 1 to relative head's right edge + 1
        final question = "Which" + answer type + continuation
    else if length of parse tree is greater than head's index + 1 AND parse [head's index + 1] is comma then
        continuation = concatenate text from parse starting at head's index + 2
        final question=answer type + "is" +continuation
    else
        reduced = question after cutting off the "For 10 ... points [name/identify]"
        final question= "Which is the" + reduced
    end if
end if
```

Original clue in QB	GPT converted question	Llama converted question
For 10 points, name this native of Rokovoko and savage companion of Ishmael in Moby-Dick.	Who is Ishmael's savage companion in Moby-Dick from Rokovoko?	What native of Rokovoko and savage companion of Ishmael in Moby-Dick?
This state's largest city endured an 1855 Rum Riot, put down with the help of 1880 Prohibitionist Candidate for President Neal S. Dow.	What state's largest city experienced the 1855 Rum Riot and was aided by Neal S. Dow in 1880?	What state's largest city endured an 1855 Rum Riot, put down with the help of 1880 Prohibitionist Candidate for President Neal S. Dow?

Table 8: Conversion of QB elicitation using two large language model-based technique (GPT and Llama). In the case of both models, we can see, the converted questions are similar.

Question	A fortification overlooking which city was renamed "narin qala" or "little fortress" by mongolinvaders in the 13th century.
Answer	Tbilisi
Positive context	City in the Caucasus, with its at least 50,000 inhabitants and thriving commerce. Several intellectuals born or living in Tbilisi, bearing the nisba al-Tiflisi were known across the Muslim world. The Abbasid Caliphate weakened after the Abbasid civil war in the 810s, and caliphal power was challenged by secessionist tendencies among peripheral rulers, including those of **Tbilisi**. At the same time, the emirate became a target of the resurgent Georgian Bagrationi dynasty who were expanding their territory from Tao-Klarjeti across Georgian lands. The Emirate of Tbilisi grew in relative strength under Ishaq ibn Isma'il, who was powerful enough to
Negative context	near the shores of Kasagh River, during the reign of king Orontes I Sakavakyats of Armenia (5702̌013560 BC). However, in his first book "Wars of Justinian", the Byzantine historian Procopius has cited to the city as "Valashabad" (Balashabad), named after king "Valash" (Balash) of Armenia. The name evolved into its later form by the shift in the medial "L" into a "Gh", which is common in the Armenian language. Movses Khorenatsi mentioned that the Town of Vardges was entirely rebuilt and fenced by king Vagharsh I to become known as "Noarakaghak" (,"New City") and later "Vagharshapat". The territory of

Table 9: We have a QB question: *A fortification overlooking which city was renamed "narin qala" or "little fortress" by mongolinvaders in the 13th century.* with answer *Tbilisi*. Now, for the positive context of the DPR training we have used those passage which contain the answer string and the rest of the passages are selected as negative context. One of the examples of positive contexts and negative contexts for this question is shown here.

Dataset	**Size**	**Wrong**	**Examples of Error**	**Comment**
Trivia QA	138384	859(0.620%)	There are around 60.000 miles of veins, arteries and capillaries in the human body. True or false? We all knew him as Radar, but was the actual first name of the pride of Ottumwa, Iowa, Corporal O'Reilly on the TV series MASH?	There are some true/false questions in TriviaQA. In our heuristics of "no wh-words", it is wrongly transformed.
Jeopardy	216930	35(0.016%)	Hits hard 1 of the 2 born in Vermont	No words to generate the question
AI King	22335	155(0.693%)	Is Ichiro a right-handed or left-handed batter in the major leagues? In horse racing, a "10,000 horse racing ticket" refers to a horse racing ticket with multiple odds? Will the 2020 Olympics in Tokyo be the Summer Olympics or the Winter Olympics?	There are some yes/no and either/or questions in the dataset. We have no heuristics to handle those clues.
Hotpot QA	90447	21(0.023%)	Are Patrick White and Katherine Anne Porter both writers? Did both Carl Boese and Franco Zeffirelli direct and produce film? Are Pam Veasey and Jon Jost both American?	There are some yes/no questions in the dataset. We have no heuristics to handle those clues.

Table 10: Error analysis of four clue-based datasets after applying our heuristics. We can see from the above analysis, is that our heuristics mostly fail to convert questions when there is an error in the question or the question is specific to the context of the game.

Algorithm 6 In rewriting elicitations into questions, we need to replace uncommon, odd answer mentions (e.g., "this polity") with more traditional ones (e.g., "this country"). Thus, we count all mentions used to refer to an answer a, then store the most frequent in M. This becomes the cannonical mention we will always use for rewriting questions. Example mentions and cannonical mentions for answers shown in Table 7.

Mention count $C := |a| \times |m|$ zero array
for Elicitation e, Answer a in Dataset **do**
    for Noun Phrase $\boldsymbol{n} \in$ Parse(e) **do**
        ▷ The mention could be any noun phrase.
        if Yield$(\boldsymbol{n})[0] \in \{$ this, these, $\ldots\}$ **then**
            ▷ Mentions start with specific determiners.
            Mention $m \leftarrow$ Yield$(\boldsymbol{n})[1:]$
            $C[a][m] \leftarrow C[a][m] + 1$
            ▷ Record all mentions of this answer
        end if
    end for
end for
Canonical Mention $M := a \mapsto m$
for Answer $a \in C$ **do**
    M[a] $\leftarrow \arg\max_m C[a][m]$
    ▷ The cannonical mention is the most frequent
end for

LLM name	Min K% probability
GLaM (Du et al., 2021a)	71.1%
FLAN (Wei et al., 2022)	62.9%
PaLM (Chowdhery et al., 2022)	68.3%
Llama (Chowdhery et al., 2022)	57.0%
T-5 (Raffel et al., 2020)	77.9%
BLOOM (Workshop et al., 2023)	64.4%
MistralOrca (OpenOrca, 2024)	47.1%
Falcon (Falcon, 2024)	55.2%

Table 11: We validate if NQ is present in their pretraining data by MIN-K(K=60)% PROB (Shi et al., 2023a). . A high average probability suggests that the NQ is likely part of the pertaining data. We can see for all the state-of-the-art LLMs, the probability is 63% on average. Thus, we can say, these models likely have NQ in their training data.

Original Question	Heuristic Applied from List in 3.1	Syntactic Transformed Question
	Dataset Name: **Jeopardy**	
For the last 8 years of his life, Galileo was under house arrest for espousing this man's theory	No wh-words	For the last 8 years of his life, Galileo was under house arrest for espousing which man's theory
The city of Yuma in this state has a record average of 4,055 hours of sunshine each year	No wh-words	The city of Yuma in which state has a record average of 4,055 hours of sunshine each year
In 1963, live on "The Art Linkletter Show", this company served its billionth burger		In 1963, live on "The Art Linkletter Show", which company served its billionth burger
Signer of the Dec. of Indep., framer of the Constitution of Mass., second President of the United States'		Who is Signer of the Dec. of Indep., framer of the Constitution of Mass., second President of the United States'
In the title of an Aesop fable, this insect shared billing with a grasshopper		In the title of an Aesop fable, which insect shared billing with a grasshopper
In the winter of 1971-72, a record 1,122 inches of snow fell at Rainier Paradise Ranger Station in this state		In the winter of 1971-72, a record 1,122 inches of snow fell at Rainier Paradise Ranger Station in which state
This housewares store was named for the packaging its merchandise came in & was first displayed on		Which housewares store was named for the packaging its merchandise came in & was first displayed on
Cows regurgitate this from the first stomach to the mouth & chew it again		Cows regurgitate this from the first stomach to the mouth & chew it again
In 1000 Rajaraja I of the Cholas battled to take this Indian Ocean island now known for its tea		In 1000 Rajaraja I of the Cholas battled to take which Indian Ocean island now known for its tea
	Dataset Name: **TriviaQA**	
Name the 1980's hit sung by Tina Turner and Rod Stewart?	Imperative to Interrogative	What is the 1980's hit sung by Tina Turner and Rod Stewart?
Name the two tiles with the highest score in Scrabble?		What is the two tiles with the highest score in Scrabble?
Name the Dick Francis mount that collapsed approaching the finishing line in the 1956 'Grand National'?		What is the Dick Francis mount that collapsed approaching the finishing line in the 1956 'Grand National'?
Name the 1972 musical starring David Essex as Jesus Christ?		What is the 1972 musical starring David Essex as Jesus Christ?
Name the male lead in the 1946 film The Big Sleep?		Who is the male lead in the 1946 film The Big Sleep?
Name the stretch of water separating Anglesey from the Welsh mainland?		What is the stretch of water separating Anglesey from the Welsh mainland?
For a point each, name the characters in a bottle of Flintstones Chewable Vitamins.		What is the characters in a bottle of Flintstones Chewable Vitamins.
For a point each, name the state(s) bordering Maine		What is the state(s) bordering Maine
Name the year: NAFTA is ratified, Nancy Kerrigan gets clubbed, Kurt Cobain eats his shotgun, OJ Simpson offs his ex wife and her friend.		What is the year: NAFTA is ratified, Nancy Kerrigan gets clubbed, Kurt Cobain eats his shotgun, OJ Simpson offs his ex wife and her friend.

Table 12: To show the generalization of our dataset, we applied the heuristics from Section 3.1 to different domain datasets. At first, heuristics are applied to two similar clue-based datasets– *Jeopardy!* and *TriviaQA*. We can see, for similar clue-like questions' datasets like QB, our heuristics convert them into NQ-like questions successfully.

Original Question	Heuristic Applied from List in 3.1	Syntactic Transformed Question
	Dataset Name: **AI King official distribution dataset**	
In 1960, while studying abroad from Nankai, he achieved a record of 5 wins, 1 loss, and 9 seasons in his one year on the job, and was promoted to the San Francisco Giants, becoming the first Japanese major leaguer.	Split Conjunction and No wh-words	In 1960, while studying abroad from Nankai, who achieved a record of 5 wins, 1 loss, and 9 seasons in his one year on the job, Who was promoted to the San Francisco Giants, becoming the first Japanese major leaguer. In 1960, while studying abroad from Nankai, who achieved a record of 5 wins, 1 loss, and 9 seasons in his one year on the job, and was promoted to the San Francisco Giants, becoming the first Japanese major leaguer.
It is Germany's second largest trading port after Hamburg, and is also featured in the Grimm fairy tales that feature musical bands.		What is Germany's second largest trading port after Hamburg, and is also featured in the Grimm fairy tales that feature musical bands? What is Germany's second largest trading port after Hamburg? What is featured in the Grimm fairy tales that feature musical bands?
This fish is said to have gotten its name from the fact that it eats by cutting its body into two?		Which fish is said to have gotten its name from the fact that it eats by cutting its body into two, but why are its ovaries called "herring roe"?
On July 16th of this year, Katsura Saegusa will become the 6th generation of the famous Kamigata Rakugo story.		On July 16th of which year, Katsura Saegusa will become the 6th generation of the famous Kamigata Rakugo story.
	Dataset Name: **Hotpot QA**	
This is the place of fish and is the capital city of Frobisher Bay south?	Split conjunction and No wh words	1. Which is the place of fish and is the capital city of Frobisher Bay south? 2. Which is the place of fish? 3. Which is the capital city of Frobisher Bay south?
This Ghanaian footballer was a notable graduate of SC Bastia Reserves and Academy?		Which Ghanaian footballer was a notable graduate of SC Bastia Reserves and Academy?
Name one comedy series that stars the younger brother of Arthur White ?		Which comedy series that stars the younger brother of Arthur White ?
Bottom Points railway station is on a heritage railway system that is situated near this town?		Bottom Points railway station is on a heritage railway system that is situated near which town?
Barry Moltz taught entrepreneurship as an adjunct professor in this city?		Barry Moltz taught entrepreneurship as an adjunct professor in which city?
Adebayo Akinfenwa was a star in the 2006 Football League Trophy Final, but know plays for this team?		Adebayo Akinfenwa was a star in the 2006 Football League Trophy Final, but know plays for which team?
Topics covered by this author include corporate control of government, the harshness of war, gender polarities and sexual identity.		Topics covered by which author include corporate control of government, the harshness of war, gender polarities and sexual identity.

Table 13: To show the generalization of our dataset, we applied the heuristics from Section 3.1 to different domain datasets. At first, heuristics are applied to to a different lingual dataset (Japanese). Secondly, it is applied to a multi-hop dataset HotpotQA. We can see, for similar clue-like questions' datasets like QB, our heuristics convert them into NQ-like questions successfully.