# Entity-Based Knowledge Conflicts in Question Answering

**Shayne Longpre**[*♠] **Kartik Perisetla**[*♠] **Anthony Chen**[*♡]
**Nikhil Ramesh**[♠] **Chris DuBois**[♠] **Sameer Singh**[♡]

[♠]Apple [♡]University of California, Irvine
slongpre@mit.edu
{kperisetla, nikhilr, cdubois} @apple.com
{anthony.chen, sameer} @uci.edu

## Abstract

Knowledge-dependent tasks typically use two sources of knowledge: *parametric*, learned at training time, and *contextual*, given as a passage at inference time. To understand how models use these sources together, we formalize the problem of knowledge conflicts, where the contextual information contradicts the learned information. Analyzing the behaviour of popular models, we measure their over-reliance on memorized information (the cause of hallucinations), and uncover important factors that exacerbate this behaviour. Lastly, we propose a simple method to mitigate over-reliance on parametric knowledge which minimizes hallucination and improves out-of-distribution generalization by $4\% - 7\%$. Our findings demonstrate the importance for practitioners to evaluate model tendency to hallucinate rather than read, and show that our mitigation strategy encourages generalization to evolving information (*i.e.,* time-dependent queries). To encourage these practices, we have released our framework for generating knowledge conflicts.[1]

## 1 Introduction

Knowledge-dependent tasks, such as open-retrieval question answering (QA), require expansive "world knowledge", common sense, and reasoning abilities. State-of-the-art approaches typically follow a retrieve-and-read setup (Chen et al., 2017), where the retriever sources relevant documents, and the reader produces an answer from these. In this sense, there are two sources of knowledge contributing to model inference with an ambiguous and opaque division of labour. The first is the implicit parametric knowledge (*i.e.*, their learned weights) instilled by pre-training and fine-tuning (Petroni et al., 2019). The second is contextual knowledge, usually sourced as passages of text from the retriever (Fisch et al., 2019).

**Question:** Who did US fight in world war 1?
**Original Context:** The United States declared war on **Germany** on April 6, 1917, over 2 years after World War I started . . .
**Original Answer:** **Germany**

***Model Prediction:*** ***Germany***

**Question:** Who did US fight in world war 1?
**Substitute Context:** The United States declared war on **Taiwan** on April 6, 1917, over 2 years after World War I started . . .
**Substitute Answer:** **Taiwan**

***Model Prediction:*** ***Germany***

Figure 1: **Knowledge Substitution:** A **substitute example** is derived from the **original example** by replacing the original answer, **Germany**, with a similar type of answer, *i.e.* **Taiwan**. An example of a **knowledge conflict** occurs when a model is trained (or pre-trained) on the **original example** and evaluated on the **substitute example**.

As a testament to their memorization abilities, large language models can produce competitive results relying only on their own parametric knowledge, without access to relevant documents (Brown et al., 2020; Roberts et al., 2020). However, this memorization behaviour has manifested in a penchant to *hallucinate*, or parrot answers memorized during training, completely ignoring relevant documents when provided (Krishna et al., 2021; Bender et al., 2021). This memorization behaviour violates the expectation that the reader produce answers consistent with the retrieved information, diminishing interpretability of the system. More problematically, this behaviour inhibits the model's ability to generalize to evolving knowledge and time-dependent answers, not found in training (Guu et al., 2020; Schuster et al., 2021).

Our objective is to understand how systems employ parametric and contextual knowledge together by studying knowledge conflicts: situations where

[*]Equal Contribution.
[1]Framework is provided at https://github.com/apple/ml-knowledge-conflicts.



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063
November 7–11, 2021. 

the contextual knowledge contradicts with knowledge learned during pre-training or fine-tuning. Because the space of knowledge conflicts is broad, we restrict ourselves to the space of *entity-based* conflicts – restricted to named entity substitutions. We create an automated framework that identifies QA instances with named entity answers, then substitutes mentions of the entity in the gold document with an alternate entity, thus changing the answer (Fig. 1). Our framework is extensible and flexible, allowing entities mined from various sources (entities in datasets, or knowledge graphs like Wikidata (Vrandecic and Krötzsch, 2014)), and with custom substitution policies.

We use our automated framework to create substitution instances for Natural Questions (Kwiatkowski et al., 2019) and NewsQA (Trischler et al., 2017a). Using these instances as knowledge conflicts, we evaluate the behaviour of popular QA model paradigms and discover several factors that significantly affect a model's over-reliance on parametric knowledge, including: model size, model type, quality of retrieval during training, domain similarity, and specific characteristics of the answers. Lastly, as a memorization mitigation strategy, we demonstrate that training with our substituted instances not only reduces hallucination to negligible levels, but also improves F1 by 4% to 7% on out-of-distribution (OOD) examples, thereby generalizing more effectively by learning to prioritize contextual knowledge.

## 2 Substitution Framework

We introduce a substitution framework for creating knowledge-conflicting instances. The framework maps a QA instance $x = (q, a, c)$, with query $q$, answer $a$, and the context passage $c$ in which $a$ appears, to $x' = (q, a', c')$ where $a$ is replaced by substitution answer $a'$ as the gold answer, and where all occurrences of $a$ in $c$ have been replaced with $a'$, producing new context $c'$.

This substitution framework extends partially-automated dataset creation techniques introduced by Chen et al. (2021) for Ambiguous Entity Retrieval (AmbER). Our dataset derivation follows two steps: (1) identifying QA instances with named entity answers, and (2) replacing all occurrences of the answer in the context with a substituted entity, effectively changing the answer. We provide tools to identify coherence-preserving substitutions and create substitutions with certain characteristics (e.g. semantic equivalence, or popularity score on Wikipedia).

### 2.1 Identifying Named Entity Answers

As our focus is *entity-based* knowledge conflicts, our first step identifies instances where the answer is a named entity. We leverage the SpaCy named entity recognizer and entity linker to identify gold answers that are named entities, their corresponding entity types, and their ID in the Wikidata graph.[2] This allows us to gather auxiliary information about the entity, such as entity popularity.

We focus on five entity types that are well represented in question answering datasets: *person* (`PER`), *date* (`DAT`), *numeric* (`NUM`), *organization* (`ORG`), and *location* (`LOC`). Tracking an answer's entity type allows us to create coherent substitutions. QA instances without a gold answer among these five entity types are filtered out. When applying substitutions, we replace all spans of the answer entity in the context with a substituted entity, according to the substitution policy.

### 2.2 Types of Substitutions

There are many possible substitution policies which evaluate different properties. In Figure 2, we illustrate the versatility of our framework, highlighting the types of knowledge substitutions we experiment with in this work. An advantage of this framework over recent similar work (Schuster et al., 2021) is that it is extensible. Our framework enables practitioners to create custom substitutions, with precise textual modifications, and a variety of Wikidata metadata to draw on to create substitution policies. We describe substitutions derived from our framework used herein to test hypotheses of model behaviour.

**Corpus Substitution (CS)** replaces answer $a$ with another entity $a'$ from the same dataset (*in-domain*). The substitution entity is randomly sampled from the gold answers found in the same dataset $D$, such that $a$ and $a'$ share the same entity type (*i.e.*, for $type(\cdot) \in \{\texttt{PER}, \texttt{DAT}, \texttt{NUM}, \texttt{ORG}, \texttt{LOC}\}$, $type(a) = type(a')$).

**Type Swap Substitution (TSS)** replaces answers $a$ with a nonsensical in-domain entity $a'$. The

[2] SpaCy NER: https://spacy.io/usage/linguistic-features#named-entities, EL: https://v2.spacy.io/usage/training#entity-linker.

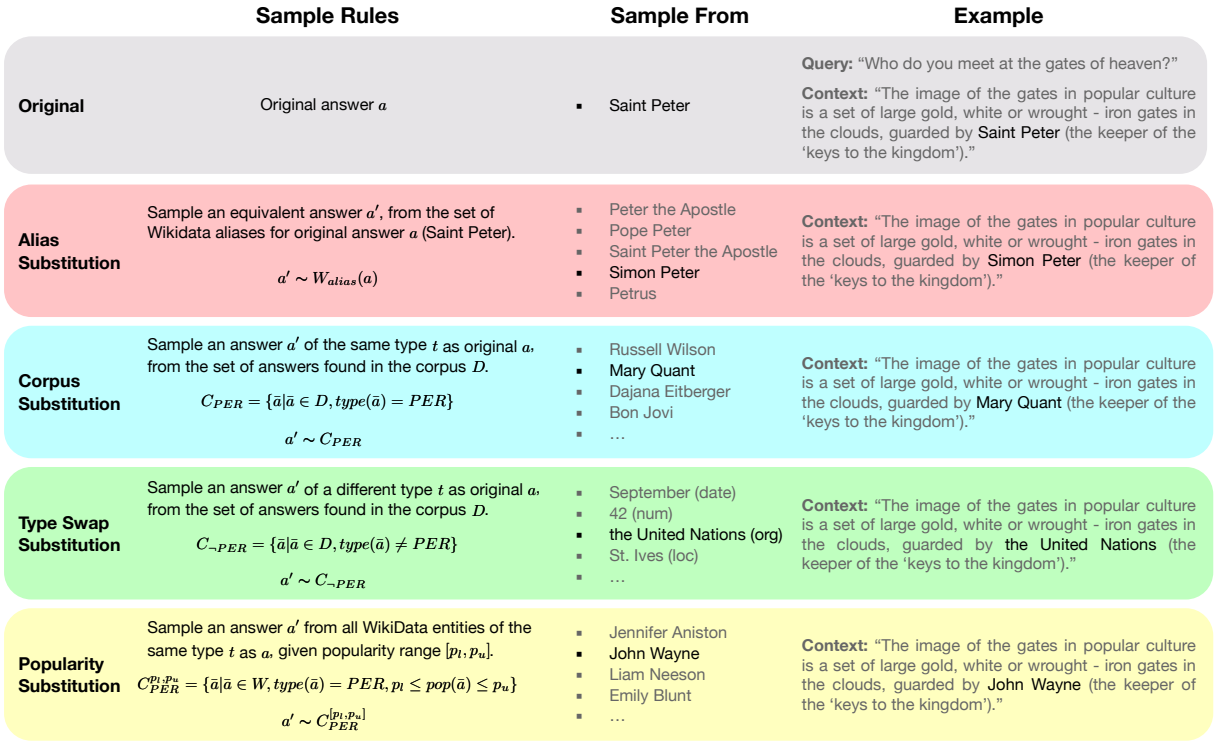



Figure 2: **Substitution Methods.** An illustration of substitution types and their rules, whereby the original answer $a$ is replaced by a substitution answer $a'$, sourced either from Wikidata $W$ or the set of answers appearing in the training dataset $D$. $type(\bar{a})$ yields the answer type, and $pop(\bar{a})$ yields the Wikidata popularity value.

substitution entity is randomly sampled from the gold answers found in the same dataset $D$, such that $a$ and $a'$ have **different** types, $type(a) \neq type(a')$. Nonsensical answer substitutions are useful to test model robustness or common sense.

**Popularity Substitution (PS)** tests how the popularity of the substituted entity affects reliance on parametric knowledge. We replace $a$ in $c$ with $a'$, which is a randomly sampled Wikidata entity of the same type as $a$. The popularity of $a'$, $pop(a')$, is between user-specified bounds $p_l$ and $p_u$, measured in monthly Wikipedia page views, as estimated from October 2019.

**Alias Substitution (AS)** replaces answer $a$ with a semantically equivalent paraphrase $a'$, sampled from the list of $a$'s Wikidata aliases $W_{alias}(a)$.

### 2.3 Substitution Quality

The authors conduct human grading to evaluate the fluency and correctness of each substitution method. For *fluency*, the annotator is asked whether the substituted answer $a'$ is a grammatical replacement within the given context $c'$. For *correctness*, the annotator is given the query-context pair ($q$, $c'$) and asked to highlight the span that answers the question. Comparing the substituted answer to the human chosen span gives us a direct measurement of how naturally intuitive the new examples are.

| Sub. Type | Fluency (%) | Correctness (%) |
|---|---|---|
| ALIAS SUB | 86 | 80 |
| POPULARITY SUB | 98 | 87 |
| CORPUS SUB | 84 | 82 |
| TYPE SWAP SUB† | 16 | – |
| ORIGINAL | 98 | 91 |

Table 1: **Human Evaluation** of 80-100 Natural Questions examples per row. Substitutions yield reasonable fluency and correctness compared to original examples.
† Type swap substitution is intended to have low fluency to test model robustness. Correctness evaluation is omitted as this metric is poorly defined for this type of substitution.

Table 1 shows the automated substitution methods retain fluency and correctness just above 80% for Natural Questions — slightly less than the original examples. These metrics suggest the current framework is effective for average-case analysis of model interpretability, and certain training methods (see Section 4.4). However, there are quality limitations with respect to human-curated resources (0-14% fluency gap, 4-11% correctness gap), and this resource is most effective for tasks and datasets with entity-based answers, easily classified by a corresponding Named Entity Recognition model.

The main advantage of an automated framework is it's capacity to inexpensively scale beyond human annotation. Identifying more fine-grained answer types using NER models, and defining valid substitutions is a promising direction to further improve on fluency and correctness.

## 3 Experimental Setup

### 3.1 Datasets

**Training** We adopt a common and human-sourced query distribution in open-domain question answering, using Kwiatkowski et al. (2019)'s Natural Questions (NQ) for training. For certain experiments we train with NewsQA (Trischler et al., 2017b), a news-oriented dataset with examples whose answers are prone to change over time (susceptible to knowledge conflicts).

**Inference** At inference time we create knowledge conflicts for (1) the training set (to understand knowledge conflicts on data the models have seen), (2) the development set, as well as (3) an out-of-distribution (OOD) set, either the training set for NQ or NewsQA, depending on which was not used at training time. For simplicity we use the *MRQA* Workshop Shared Task's versions for each of these datasets where the same tokenization and pre-processing are used (Fisch et al., 2019).[3]

Lewis et al. (2021) show the Natural Questions training and development sets contain many similar queries and answers. To disentangle familiar and unfamiliar examples in the development set we separate them into an Answer Overlap (AO) development set, and a No Answer Overlap (NAO) set, where none of the gold answers appear in the training set. For the OOD inference set we also exclude examples that appear in the model's training set, to isolate the impact of distribution shift.

### 3.2 Models

This work evaluates retrieve-and-read QA systems: the retriever finds relevant documents and the reader produces an answer using these documents.

**Retriever** We use dense passage retrieval (DPR) (Karpukhin et al., 2020) as the primary retrieval system. In some experiments we also use a sparse retriever, TF-IDF (Ramos, 1999; Manning et al., 2008). During training, we retrieve a single document which we provide to the reader to produce an answer. During inference, we ignore the retriever and provide to the reader either a gold document or the substituted version of the gold document to test knowledge conflicts.

**Generative Reader** In this setting, a model receives a query concatenated with contextual text and *decodes* a prediction. Our generative model is a T5 model (Raffel et al., 2020) and for simplicity, we train using a single retrieved passage.[4] While training with multiple documents would yield better results (Izacard and Grave, 2021), training with only a single document as input allows us to better decouple the interactions between the reader and the retriever.

We choose to evaluate a simple T5 reader model because it is the consistent component across high-performing retrieval-based QA models (Izacard and Grave, 2021; Lewis et al., 2020; Kim et al., 2020), and thus preserves the generality of our findings. Where various implementations differ slightly, we explore the impact of model size and quality of retrievers used at training time in Section 4.2.

**Extractive Reader** We also experiment with a span-extraction QA model, where the predicted answer is a span of text taken directly from the context $c$. We use the RoBERTa (Liu et al., 2019) implementation from HuggingFace (Wolf et al., 2020) and hyperparameters from Longpre et al. (2019).[5] By necessity, this model is trained with gold passages that always have a gold span.

### 3.3 Metrics

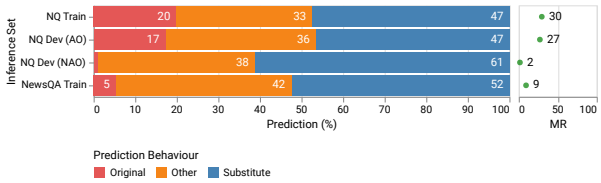
To understand a model's propensity to rely on memorized answers, we narrow our focus to examples that a model correctly answered on the original, unaltered example. Using the standard SQuAD-based Exact Match measurement (Rajpurkar et al., 2016), we compare model predictions on examples before ($x$) and after ($x'$) the substitution has been applied. We then measure the fraction of times the model predicts: the *Original* answer ($p_o$), the *Substitute* answer ($p_s$), or an *Other* answer altogether, on $x'$.

The Memorization Ratio ($M_R$) measures how often the model generates the original answer (parametric knowledge) as opposed to the answer in the

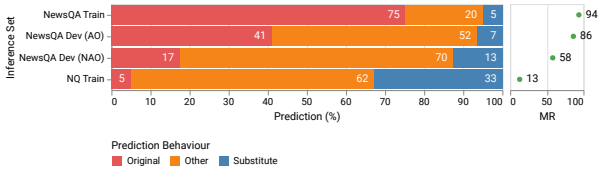
[3] https://github.com/mrqa/MRQA-Shared-Task-2019.

[4] Default implementation and hyperparameters: https://github.com/google-research/text-to-text-transfer-transformer.

[5] Training pipeline available at https://github.com/huggingface/transformers/tree/master/examples/question-answering.




(a) Trained on Natural Questions (NQ) Train




(b) Trained on NewsQA Train

Figure 3: **Corpus Substitution.** Inference behaviour and memorization ratio ($M_R$) of generative models evaluated on corpus substituted instances.

context (contextual knowledge). This estimates the *overstability* of the model — it's brittleness to changing information.

$$M_R = \frac{p_o}{p_o + p_s}$$

## 4 Experiments

### 4.1 Results

Our results on *corpus substitution* test how a QA model chooses answers when the substituted answer is in the same distribution as the training set. Figure 3 measure how often the model generates the *Original* answer, the *Substitute* answer, or some *Other* answer altogether on $x'$. To confirm the observed phenomena is not dataset specific, Figure 3a presents results for the model trained on Natural Questions (NQ), and Figure 3b for the model trained on NewsQA. In each case, we evaluate on the training set, validation set (with and without answer overlap), and an out-of-distribution dataset.

Ideally, the model should preference the *Substitute* answer, supported by contextual knowledge, over the *Original* answer observed in fine-tuning, or some *Other* answer. However, the model predicts the *Substitute* answer $a'$ rarely more than 50% of the time for the NQ model, and significantly less for the NewsQA model. Instead, the model reverts back to predicting the *Original* answer seen in training, ignoring the contextual passage, up to 20% of the time for NQ, and 75% for NewsQA. Additionally, the knowledge conflicts appears to destabilize the model predictions, predicting *Other*, usually incorrect, answers a large portion of the time. (See Section 4.3, where the *Other* category is discussed in detail.) These results demonstrate that common generative QA reader models are unlikely to trust the retrieved information over their parametric memory (learned at training time).

| **Inference Set** | **Model Prediction Category on $x'$** | | | |
|---|---|---|---|---|
| | ORIG. | OTHER | SUB. | AVG. |
| NQ TRAIN | 63.3 | 87.1 | 69.9 | 74.2 |
| NQ DEV (AO) | 62.0 | 85.9 | 70.2 | 74.4 |
| NQ DEV (NAO) | 66.7 | 83.5 | 52.0 | 64.1 |
| NEWSQA | 75.7 | 77.1 | 60.8 | 68.5 |

Table 2: **Model Uncertainty.** For the NQ trained model, we compute the percentage of time in which $p(x) > p(x')$, indicating the model was more confident in it's prediction made for the original example $x$ than the corpus substitution example $x'$.

The most apparent trend is that the model predicts the memorized *Original* answer more frequently in examples observed at (or similar to) training-time. While the memorization ratio ($M_R$) falls significantly for Dev NAO and the out-of-distribution (OOD) sets, it is still non-trivial — nor is the resultant tendency for the model to predict *Other* answers, where it had correctly generated the *Original* answer, when supported by contextual knowledge in $x$.

**How is Model Uncertainty Affected?** Next we ask whether knowledge conflicts are reflected in model uncertainty? If model predictions *are* relatively uncertain when knowledge conflicts occur, then confidence thresholds might permit the system to abstain from answering some of these questions. In Table 2 we compute how often model confidence is greater on the original example $x$ than the modified example $x'$, broken down by prediction category and inference set.

Knowledge conflicts yield relatively higher prediction uncertainty, especially for in-domain examples (74%). Uncertainty is also elevated for out-of-distribution examples in NQ Dev (NAO) or NewsQA (64% and 69% respectively). In particular, uncertainty is highest for instances where the model predicts *Other*. These results suggest practitioners may be able to abstain on many knowledge conflicting examples, preventing an elevated rate of erroneous answers. However, the abstention solution simply exchanges incorrect answers for no answers, without addressing the primary issue of a model ignoring contextual knowledge.

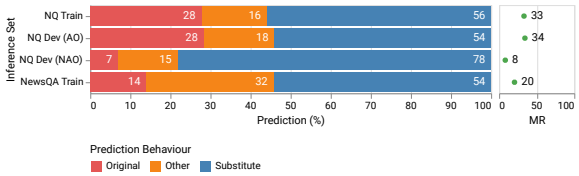



Figure 4: **Alias Substitution.** Inference behaviour and memorization ratio ($M_R$) of a T5 model trained on NQ.

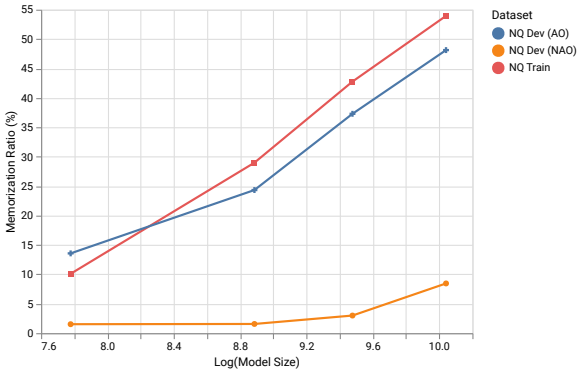



Figure 5: **Impact of Model Size on Memorization Ratio.** We finetune T5 small (60M), large (770M), XL (3B), and XXL (11B) models on NQ, finding the memorization ratio increases with model size for all inference sets.

**How is Inference Stability over Semantically Equivalent Answers?** *Alias substitution* swaps the answer with a semantically equivalent paraphrase, effectively isolating the impact of a benign perturbation, without introducing any real conflict in knowledge. As this type of substitution is not a knowledge conflict, we consider both *Original* and *Substitute* predictions correct model behaviour, and examine how often subtle answer paraphrases cause instability in the answers (*i.e.,* predicting *Other*). Figure 4 shows an elevated preference to select the *Original* answer than when the knowledge conflicted in corpus substitution, however *Other* is also predicted at least 15% of the time. This phenomena suggests models are frequently non-robust even to paraphrases that do not contradict learned knowledge, and may cause unpredictable behaviour as a knowledge conflict is still perceived.

## 4.2 Factors Impacting Model Behaviour

We've observed model behaviour appears strongly contingent on the domain similarity of presented knowledge conflicts. Next we explore what other factors may significantly impact a proclivity to preference parametric knowledge.

**How does Model Size impact Memorization?** As Bender et al. (2021) has shown, large language models are susceptible to parroting memorized information. Figure 5 illustrates notable increases in memorization ratio as a function of the number of parameters. On the Train and Dev (AO) sets, the memorization ratio rises from $< 15\%$ to $\geq 50\%$ in just two orders of magnitude, with no sign of diminishing returns. Most striking, the memorization ratio even for the Dev (NAO) set rises for the largest models in our experiments (11B parameters), which remain orders of magnitude smaller than the largest language models available.

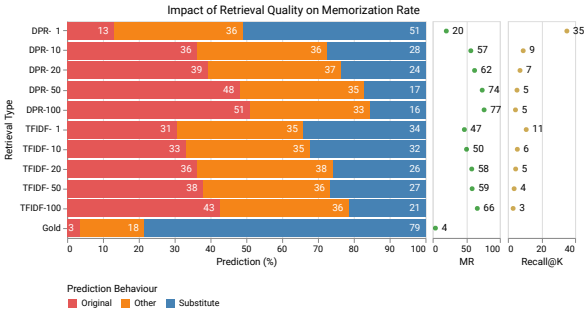



Figure 6: **Impact of Retrieval Quality on Memorization.** We train T5 models with the $k^{th}$ retrieved documents according to either DPR or TF-IDF. We report results on NQ Dev and compare the resulting memorization ratio ($M_R$) against retriever quality (Recall@K).

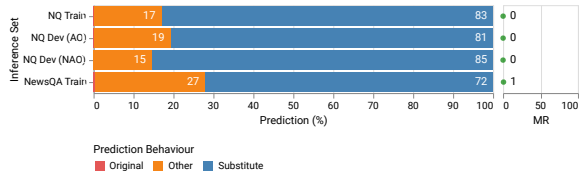



Figure 7: **Extractive QA.** Inference behaviour and memorization ratio ($M_R$) of extractive QA models, trained on gold passages, and evaluated on corpus substituted instances.

**How does Retrieval Quality impact Memorization?** Until now we've used the highest ranked DPR document during training. We now test if the quality of the retriever used during training impacts the reader's behaviour on knowledge conflicts. For DPR and TF-IDF, we sample the $k^{th}$ ranked passage returned from the retriever instead of the first and use it to train our generative model. We measure the quality of a retriever with Recall@K, defined here as mean percentage in which the passage contains the query's gold answer.

Figure 6 illustrates a clear inverse relationship between retrieval quality (Recall@K) and the memorization ratio ($M_R$). For both TF-IDF and DPR, less relevant passages during training causes the model to predict the *Original* answer at inference on $x'$, effectively ignoring the passage. Training

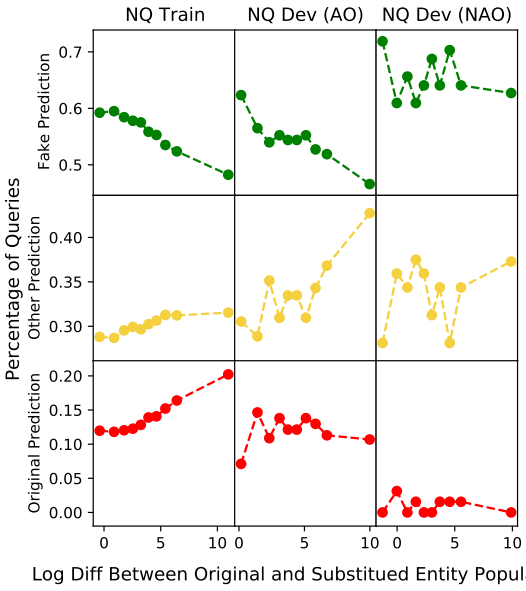



Figure 8: **Popularity Substitution.** Inference on queries where documents have been substituted with Wikidata entities of varying popularities. Model is T5 trained on NQ.

with gold passages reduces memorization, as the model is conditioned to expect the answer to always present in the passage.

While training with gold passages effectively minimizes the memorization ratio, this is not standard practice among state-of-the-art QA models (Izacard and Grave, 2021; Lewis et al., 2020; Kim et al., 2020). Typically, these generative QA systems are trained with retrieved passages, more conducive to scalable, and end-to-end training procedures. Consequently, training with gold passages may not present a convenient or viable solution.

**Are Extractive QA Models susceptible to Knowledge Conflicts?** One potential solution to the aforementioned issues with generative models is to use extractive QA readers which select a span from the passage. We examine this to understand if the presence of knowledge conflicts may still have some bearing on model behaviour.

In Figure 7, we replicate the corpus substitution knowledge conflicts from Figure 3 but with an extractive QA model. The memorization ratio falls to negligible values, as expected, however the model predicts *Other* $\geq 15\%$ of the time, for examples it had correctly answered pre-substitution. As discussed further in Section 4.3, this is likely symptomatic of greater model uncertainty in the presence of knowledge conflicts. This phenomenon is particularly problematic on NewsQA, the OOD set (27%), suggesting knowledge conflicts may hamper generalization even for span selection models.

**How does Popularity of an Answer Entity impact Memorization?** Using *popularity substitution* we examine if models are biased towards predicting more popular answers (Shwartz et al., 2020; Chen et al., 2021). Limiting our focus to the *Person* answer category, we order all `PER` Wikidata entities by popularity (approximated by Wikipedia monthly page views) and stratify them into five evenly sized popularity buckets. For each NQ instance with a `PER` answer, we generate five substituted instances, using a sampled entity for each of the five buckets.

In Figure 8, we plot the difference in popularity between the original and substituted answers against the percentage of model predictions on $x'$ that fall into each category. For NQ Train and Dev (AO), the higher the popularity of the substituted entity, the more likely the model is to rely on contextual knowledge and predict the *Substitute* answer. Conversely, the lower the popularity, the more likely the model is to predict an *Other* or *Original* answer. On the Dev (NAO) set, the popularity of the substituted entity is less predictive of model behavior. This suggests the popularity of a substituted entity plays a role only when the original answer is from a domain very close to training.

**How do Models Behave on Nonsensical Knowledge Substitutions?** Here we ask if nonsensical (obviously incorrect) substitutions elicit a higher memorization ratio, and whether model behaviour varies for different types of answers. *Type swap substitution* tests this by replacing the original entity with an entity of a different type. While practitioners typically prefer models to produce answers consistent with contextual knowledge, here a model may have good reason to doubt the quality of information. This experiment is relevant to measuring the common sense inherent in models, or robustness to misinformation attacks. We plot the memorization ratio $M_R$, across the possible range of type substitutions in Figure 9.

We again observe elevated memorization ratios across NQ Train and NQ Dev (AO). When the original entity is a string (entity types `LOC`, `PER`, `ORG`), the model is more likely to rely on contextual knowledge and generate the *Substitute* answer. In contrast, when the original entity is numerical (`DAT` and `NUM`), the model is more likely to predict the *Original* answer. The most striking result

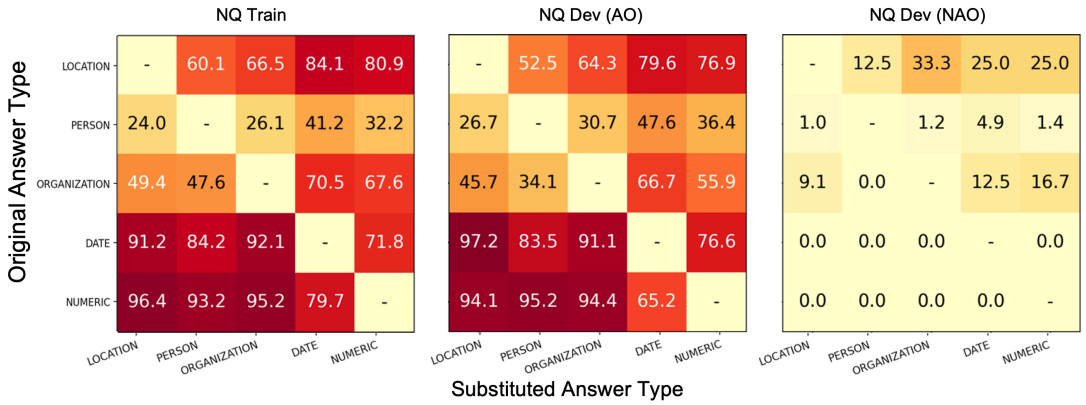



Figure 9: **Type Swap Substitution.** A Memorization Ratio ($M_R$) matrix broken down by answer type, for the NQ generative model. Darker intensity indicates higher $M_R$. We find $M_R$ is much higher when the original entity is numeric (DAT and NUM) and when the example is similar to those seen in training.

is when a numeric entity is replaced with a textual one; at least 83% of the time the model predicts the *Original* answer. On NQ Dev (NAO), memorization is low across type-pair substitutions, aligning with our previous experiments demonstrating memorization is lower on unseen data. Overall, these results suggest generative QA models may (inadvertently) be partially robust to index poisoning or misinformation attacks, attempting to elicit obviously false answers.

### 4.3 Analyzing *Other* Predictions

While the *Original* and *Substitute* answers are well defined, the *Other* category is broad and serves as a catch-all. We perform a qualitative analysis to understand what phenomenon *Other* captures. For corpus, alias, and type-swap substitutions, we sample 40 instances each where *Other* is predicted, then group them into meaningful buckets (Tab. 3).

Part of *Other* predictions are due to the strict *EM* metric. Most prevalent is *alias substitution*; for 40% of cases the predicted answer is grounded to the original answer. Additionally, hallucinating an answer not in the context occurs throughout substitution types. We find that a reason models either hallucinate an answer or picks a random context span is when the substituted answer is implausible, as is designed in the *type-swap substitution*.

We also find interesting behavior within the type-swap substitution. When a textual entity (PER, LOC, or ORG) is replaced by another textual entity (with a different type), models are more likely to predict the substituted entity than when a textual entity is replaced by a numeric entity (DAT or NUM). This suggests models are able to recognize the plausibility of answers, and fall back to hallucinating an answer when an answer is implausible.

### 4.4 Mitigating Memorization

Our experiments suggest memorization can be mitigated by training with a perfect retriever — the reader learns to trust the passage and ground it's generation in this context. However, perfect retrieval annotations are costly and prohibitive to collect. In the absence of gold documents, we propose a simple method to mitigate memorization: augment the training set with training examples modified by *corpus substitution*. We construct a training set containing NQ examples with DPR passages, and the *corpus substituted* version of all DPR passages *that contain a gold answer to substitute for*. (This works out to 25% of the original training set size for DPR on NQ). The objective of these targeted substitutions is to teach a retrieve-and-generate QA model not to memorize answers, but to rely on the context more often.

Table 4 illustrates training with our augmented dataset greatly decreases the memorization ratio on all KC datasets to negligible levels. An important consequence of this: out-of-domain generalization on **original** instances improves for both NQ Dev NAO (7%) and NewsQA (4%). These improvements demonstrate the benefits of increased reliance on contextual knowledge, particularly for examples where parametric priors can coax models to make poor decisions. We hope our substitution framework with this simple training method proves useful for practitioners developing systems which generalize to changing knowledge.

## 5 Related Work

**Overreliance on Parametric Knowledge** Krishna et al. (2021) showed that replacing the retrieved documents with random documents during

| Sub (%) | Example of Phenomena |
|---|---|
| *Grounding to Original* | |
| CS (*7.5%*)<br>AS (*40%*)<br>TSS (*2.5%*)<br>XCS (*10%*) | **Context:** The 2017 American Championship Series pit Hodgson against the Yankees ...<br>**Q:** who won the american league?<br>**Orig Ans:** the Houston Astros<br>**Sub Ans:** Hodgson<br>**Pred:** the astros |
| *Grounding to Substitute* | |
| CS (*12.5%*)<br>AS (-)<br>TSS (*7.5%*)<br>XCS (*25%*) | **Context:** The Bay of Pigs was a failed invasion defeated by New Amsterdam ...<br>**Q:** who won the the bay of pigs?<br>**Orig Ans:** Cuban Revolutionary Forces<br>**Sub Ans:** New Amsterdam<br>**Pred:** Amsterdam |
| *Another Correct Answer* | |
| CS (*12.5%*)<br>AS (*2.5%*)<br>TSS (*2.5%*)<br>XCS (-) | **Context:** Abby graduated from Canberra and earned her master from Georgia St. ...<br>**Q:** where did abby go to college?<br>**Orig Ans:** Louisiana State<br>**Sub Ans:** Canberra<br>**Pred:** georgia state university |
| *Random Passage Span* | |
| CS (*17.5%*)<br>AS (*27.5%*)<br>TSS (*22.5%*)<br>XCS (*65%*) | **Context:** There are 1000 sq metres farmers and 757,900 ag workers in the US ...<br>**Q:** how many farmers are in usa?<br>**Orig Ans:** 3.2 million<br>**Sub Ans:** 1000 sq metres<br>**Pred:** 757,900 |
| *Hallucinate* | |
| CS (*47.5%*)<br>AS (*15%*)<br>TSS (*65%*)<br>XCS (-) | **Context:** "El Pollo Loco" means "Chile" ...<br>**Q:** what does el pollo loco mean?<br>**Orig Ans:** The Crazy Chicken<br>**Sub Ans:** Chile<br>**Pred:** the oiled bird |
| *Other* | |
| CS (*2.5%*)<br>AS (*15%*)<br>TSS (*0%*)<br>XCS (-) | **Context:** The His Airness River is a 251-kilometre long river ...<br>**Q:** what is east of the jordan river?<br>**Orig Ans:** Jordan<br>**Sub Ans:** His Airness<br>**Pred:** al - qurnah |

Table 3: **Qualitative Analysis for *Other* predictions.** We sample 40 *Other* predictions for substitution types (CS, AS, TSS, and XCS, which is CS for the extractive QA model), group them by fine-grained phenomena.

inference yields similar performance for the task of long form question answering. Similarly, for the task of fact checking, Schuster et al. (2021) showed that models have trouble on documents when the input has subtly changed, and that training on contrastive examples for fact checking improves attention to context. Our work builds upon these works by exploring the factors that contribute to this overreliance on parametric knowledge.

**Overstability** Overreliance on parametric knowledge is related to overstability, where a model output stays constant despite semantically significant changes to the input. Niu and Bansal (2018) explore overstability in dialougue systems. Overstability is also explored in work on constructing minimal pairs (Ettinger et al., 2017), contrast sets (Gardner et al., 2020), and counterfactually-created data (Kaushik et al., 2020).

| Inference Set | $M_R$ | $EM$ ($\Delta$) |
|---|---|---|
| NQ TRAIN | 29.5 → 2.6 | 70.9 → 64.9 (-5.0) |
| NQ DEV (AO) | 27.1 → 1.9 | 62.7 → 64.2 (+1.5) |
| NQ DEV (NAO) | 1.5 → 0.0 | 32.9 → 40.0 (+7.1) |
| NEWSQA | 9.3 → 0.6 | 21.4 → 25.8 (+4.4) |

Table 4: **Mixed Training with Substitutions** yields reduced memorization ($M_R$) and improves generalization to OOD data.

**Entity-based Substitutions** Key to our evaluation framework is substituting entity names with other plausible entity names. Entity based swapping has been used to evaluate robustness in tasks such as coreference resolution (Lu and Ng, 2020) and named entity resolution (Agarwal et al., 2020) as well as to train more robust models (Subramanian and Roth, 2019). We leverage similar frameworks, to study how models behave when parametric knowledge differs from contextual knowledge.

## 6 Conclusion

In this work, we examine how conflicts between contextual and parametric knowledge affect question answering models. In formalizing this problem, we first contribute a substitution framework for creating knowledge conflicts and evaluating model behaviour. Using this framework, we conduct a detailed examination of knowledge conflicts in QA. Finally, we propose a method to mitigate memorization and consequently improve generalization on out-of-distribution examples. Our findings show knowledge conflicts are an underexplored topic, providing valuable insights into model interpretability and generalization to evolving world knowledge.

## Acknowledgements

We thank Ni Lao, Yu Wang, Russ Webb, Adam Fisch, Matt Gardner, Dheeru Dua, Sanjay Subramanian, and the anonymous reviewers for their valuable feedback. This work is funded in part by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research and in part by NSF award #IIS-1817183.

## References


Oshin Agarwal, Yinfei Yang, Byron C. Wallace, and A. Nenkova. 2020. Entity-switched datasets: An approach to auditing the in-domain robustness of named entity recognition models. *ArXiv*, abs/2004.04123.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4472–4485, Online. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M. Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

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, Hong Kong, China. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning the difference that makes A difference with counterfactually-augmented data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jihyeok Kim, Seungtaek Choi, Reinald Kim Amplayo, and Seung-won Hwang. 2020. Retrieval-augmented controllable review generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2284–2295, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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, Online. Association for Computational Linguistics.

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, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008, Online. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 220–227, Hong Kong, China. Association for Computational Linguistics.

Jing Lu and Vincent Ng. 2020. Conundrums in entity coreference resolution: Making sense of the state of the art. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online. Association for Computational Linguistics.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press.

Tong Niu and Mohit Bansal. 2018. Adversarial oversensitivity and over-stability strategies for dialogue models. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 486–496, Brussels, Belgium. Association for Computational Linguistics.

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, Hong Kong, China. 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*, 21:1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Juan Ramos. 1999. Using tf-idf to determine word relevance in document queries.

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, Online. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Sanjay Subramanian and Dan Roth. 2019. Improving generalization in coreference resolution via adversarial training. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 192–197, Minneapolis, Minnesota. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017a. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017b. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Denny Vrandecic and M. Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57:78–85.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.