Exploring The Landscape of Distributional Robustness for Question Answering Models

Anas Awadalla[1] Mitchell Wortsman[1] Gabriel Ilharco[1] Sewon Min[1]

Ian Magnusson[2] Hannaneh Hajishirzi[1,2] Ludwig Schmidt[1,2]

Abstract

We conduct a large empirical evaluation to investigate the landscape of distributional robustness in question answering. Our investigation spans over 350 models and 16 question answering datasets, including a diverse set of architectures, model sizes, and adaptation methods (e.g., fine-tuning, adapter tuning, in-context learning, etc.). We find that, in many cases, model variations do not affect robustness and in-distribution performance alone determines out-of-distribution performance. Moreover, our findings indicate that i) zero-shot and in-context learning methods are more robust to distribution shifts than fully fine-tuned models; ii) few-shot prompt fine-tuned models exhibit better robustness than few-shot fine-tuned span prediction models; iii) parameter-efficient and robustness enhancing training methods provide no significant robustness improvements. In addition, we publicly release all evaluations to encourage researchers to further analyze robustness trends for question answering models.

1 Introduction

Over the past few years, natural language processing has seen substantial progress. In many benchmarks, large pre-trained models adapted to a target dataset reach or even surpass human performance (Devlin et al., 2019; Raffel et al., 2019; Radford et al., 2019; Brown et al., 2020b; Hoffmann et al., 2022; Chowdhery et al., 2022, inter alia). At the same time, current methods still fail to generalize reliably in a variety of test conditions (Ribeiro et al., 2020; Gardner et al., 2020; Koh et al., 2021; Luu et al., 2021; Ribeiro and Lundberg, 2022), which limits their applicability and raises questions about what exactly the methods learn (Bender and Koller, 2020). One limitation of current benchmarks is that they often measure performance only on data that comes from the same distribution as the training set (Wang et al., 2018, 2019a). However, evaluating models on a single test set provides no information on whether a method also performs well under distribution shift. While there is an increasing amount of research on robustness in NLP (Ribeiro et al., 2020; Tu et al., 2020; Hendrycks et al., 2020; Gardner et al., 2020; Arora et al., 2021; Veitch et al., 2021; Goel et al., 2021; Miller et al., 2020, inter alia), the community has not yet adopted a common set of best practices for evaluating robustness. As a result, new methods often do not evaluate on comparable or even any robustness test sets, which makes it challenging to understand which methods generalize more reliably and whether NLP is making progress on robustness to distribution shift.

To address this challenge and shed light on the robustness landscape in NLP, we conduct a large empirical evaluation of distributional robustness in question answering (QA). Building on recent research on robustness in computer vision (Taori et al., 2020; Miller et al., 2021a), we focus on distribution shifts that arise between two related but different test sets. These distribution shifts are sometimes called dataset shift to distinguish them from other kinds of distribution shift. An example of dataset shift is a pair of QA test sets where one test set is constructed from Wikipedia articles and the other from Amazon product reviews, possibly also with a different crowdsourcing process. In contrast to other notions of robustness such as adversarial robustness, dataset shifts involve no synthetic perturbations of existing test examples and are therefore more representative of generalization challenges arising "in the wild" (Taori et al., 2020).

Within the scope of dataset shifts for QA, our robustness evaluation includes a wide range of models and distribution shifts. Specifically, we assembled a testbed of over 350 QA models and 16 QA datasets, including SQuAD v1.1 (Rajpurkar et al., 2016), SquadShifts (Miller et al., 2020), and

[1]University of Washington. [2]Allen Institute for AI. Correspondence to anasa2@cs.washington.edu.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5971–5987
December 7-11, 2022 ©2022 Association for Computational Linguistics

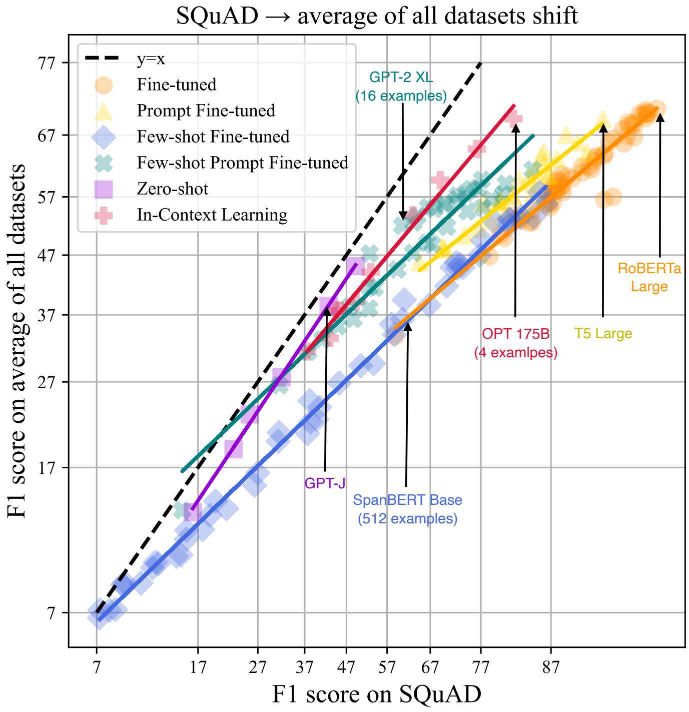


Figure 1: We evaluate over 350 models on 16 datasets to characterize the landscape of distributional robustness in question answering. Our results span a variety of architectures and adaptation strategies, including zero-shot inference, fine-tuning, and in-context learning (ICL). The x-axis shows performance on SQuAD (in-distribution), while the y-axis shows the average performance on the 15 other QA datasets (out-of-distribution). Almost all models lie under the $y = x$ diagonal, i.e., performance drops under distribution shift. Moreover, within certain groups of models—for instance, ICL models—in-distribution performance accurately predicts out-of-distribution performance. As in Taori et al. (2020), we apply logit axis scaling to clarify that the relationship between in-distribution and out-of-distribution performance is approximately linear in the logit domain.

MRQA test sets (Fisch et al., 2019). Our testbed spans different model architectures, model sizes, and pre-training setups. In addition, we evaluate a variety of approaches for applying pre-trained models to question answering including supervised fine-tuning, in-context learning, parameter-efficient fine-tuning, zero-shot inference, and more. Finally, we also include methods specifically designed to enhance robustness such as RXF (Aghajanyan et al., 2021) and FreeLB (Zhu et al., 2020).

Our testbed enables us to both identify overarching trends spanning many models, and to contextualize the robustness behavior of individual models. Among our findings are the following key results:

- Dataset shift still is an unsolved problem in QA: most models suffer a large performance drop under this kind of distribution shift.
- Despite different architectures and model sizes, many models follow a consistent trend relating in-distribution and out-of-distribution performance. Improving in-distribution performance usually also increases out-of-distribution performance in a predictable way.
- Current robustness interventions follow the same trend as models without such interventions, i.e., the robustness interventions do not increase robustness to dataset shifts.
- The only exception to the otherwise universal performance trend are zero-shot, in-context learning, and few-shot prompt fine-tuned models. These models are more robust than the baseline given by the other models in our testbed. However, the robustness of large decoder-only models decreases as the models are fine-tuned on more data from the target task.

Figure 1 summarizes our findings and shows the average F1 score on all distribution shifts as a function of the F1 score on SQuAD. Interestingly, our overall results are analogous to similar large-scale robustness evaluations in computer vision (Taori et al., 2020; Miller et al., 2021a; Radford et al., 2021), which suggests that there may be a shared underlying mechanism behind these distribution shifts that warrants further investigation.

We hope that our work helps clarify the state of robustness in NLP and provides a starting point for future work. To simplify measuring robustness to dataset shift and enable future robustness improvements, we will release our testbed including all 350+ models and evaluation results.

The remainder of the paper is organized as follows: first, we detail background and experimental setup (§2). Next, we introduce and answer our specific research questions (§3, 4). Finally, we discuss the limitations of our approach, overall conclusions, and directions for future investigation (§6, 8).

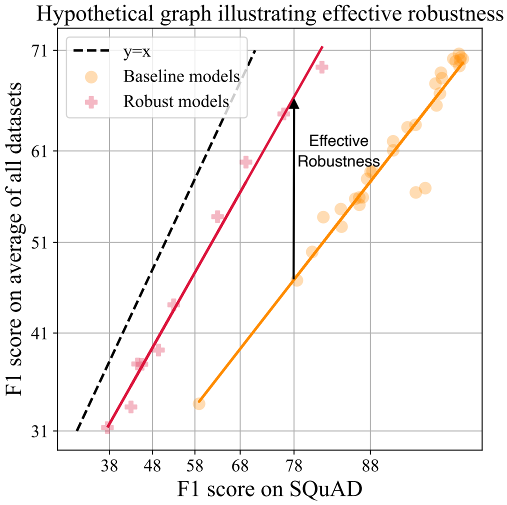


Figure 2: A schematic which illustrates the robustness measuring technique we use. *Effective robustness scatter plots* (Recht et al., 2019; Taori et al., 2020) display performance on the distribution from which training data is from (in-distribution) on the x-axis, and out-of-distribution performance on the y-axis. Effective robustness is vertical movement towards the $y = x$ diagonal beyond the baseline trend fit to fully fine-tuned models—a model with higher effective robustness has more consistent performance in- and out-of-distribution.

2 Experimental Setup

Our testbed includes over 350 models, covering a broad range of model architectures, pre-training datasets, and adaptation strategies. We use SQuAD v1.1 (Rajpurkar et al., 2016) as our reference point for question answering performance because SQuAD is a popular dataset and the performance ceiling is comparatively well understood since humans can achieve an F1 score around 95 (Miller et al., 2020). For all models except those performing zero-shot inference, we adapt the models to question answering with the SQuAD training set.

We evaluate robustness to distribution shift on the remaining 15 question answering datasets (Table 1). We follow Taori et al. (2020) in defining robustness, i.e., we say a model is robust if it has consistent performance under a distribution shift from a reference distribution to another distribution. We refer to SQuAD as in-distribution (ID) and the other 15 datasets as out-of-distribution (OOD). In the remainder of this section, we describe the different models, adaptation strategies, datasets, and evaluation details.

2.1 Models

Our testbed focuses on transformer models ranging from 11 million to 175 billion parameters. We explore several encoder-only models—ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019), SpanBERT (Joshi et al., 2020), RoBERTa (Liu et al., 2019), and Splinter (Ram et al., 2021a)—encoder-decoder models —T5 (Raffel et al., 2019) and BART (Lewis et al., 2020)—and decoder-only models (GPT-2 (Radford et al., 2019), OPT (Zhang et al., 2022), GPT-Neo (Black et al., 2021), and GPT-J (Wang and Komatsuzaki, 2021)).

2.2 Adaptation strategies

We evaluate multiple adaptation strategies—methods that *adapt* the pre-trained language model to perform better on a downstream task using labeled, in-distribution training data, e.g., through gradient based learning and in-context learning. We also examine models evaluated in a zero-shot setting, which we also refer to as an adaption method for consistency, even though no data from the in-distribution dataset is observed. For a subset of these models we also explore few-shot instead of full-shot adaptation to assess the impact of the number of training examples on robustness.

2.2.1 Fine-tuning (baseline)

We include a common fine-tuning method: adding a span prediction head and updating all the parameters in a language model via additional training on a downstream dataset, as done in Devlin et al. (2019) and subsequent work.

2.2.2 Prompt fine-tuning

Prompt fine-tuning adds no additional task specific layers and fine-tunes the existing weights to generate the answer. We use next token prediction when fine-tuning auto-regressive models like GPT. For T5 and BART models we use two fine-tuning tasks: 1) casting QA as an infilling task and generate the answer by predicting a masked span 2) conditioning the model on the context and question and fine-tune it to generate the answer.

2.2.3 Parameter-efficient fine-tuning

Parameter-efficient fine-tuning modifies only a small percentage of existing or auxiliary parameters, while freezing all other parameters. We evaluate Houlsby (Houlsby et al., 2019) and Pfeiffer (Pfeiffer et al., 2021) adapters, prefix tuning (Li and Liang, 2021), and LoRA (Hu et al., 2021). While these methods modify only a small number of parameters, they have been shown to be com-

Dataset name	Test set size	Domains
SQuAD v1.1 dev. set (Rajpurkar et al., 2016)	10,570	Wikipedia
SquadShifts New-Wiki (Miller et al., 2020)	7,938	Wikipedia
SquadShifts Reddit (Miller et al., 2020)	9,803	Reddit
SquadShifts NYT (Miller et al., 2020)	10,065	New York Times
SquadShifts Amazon (Miller et al., 2020)	9,885	Amazon reviews
RACE (Lai et al., 2017)	674	English exams from China
DROP (Dua et al., 2019)	1,503	Wikipedia
NewsQA (Trischler et al., 2017)	4,212	CNN articles
SearchQA (Dunn et al., 2017)	16,980	Jeopardy! questions with contexts from Google search
NaturalQuestions (Kwiatkowski et al., 2019)	12,836	Google search questions with contexts from Wikipedia
DuoRC (ParaphraseRC) (Saha et al., 2018)	1,501	Movie plots from IMDB and Wikipedia
HotpotQA (Yang et al., 2018)	5,904	Wikipedia
TextbookQA (Kembhavi et al., 2017)	1,503	Middle school science questions from textbooks
TriviaQA (Joshi et al., 2017)	7,785	Trivia questions with contexts collected using a Bing search
RelationExtraction (Levy et al., 2017)	2,948	Generated samples using a knowledge base
BioASQ (Tsatsaronis et al., 2015)	1,504	Medical articles

Table 1: Question answering datasets used to evaluate models in this work. SQuAD is used as the in-distribution reference dataset—we use training data from SQuAD to adapt models. The remaining datasets are used to answer the question of how SQuAD models perform under dataset shift—we use these other datasets for evaluation only.

petitive with full fine-tuning when measuring in-distribution performance. Previous work suggests freezing a majority of model weights may make these methods more robust (Lester et al., 2021).

2.2.4 Robustness enhancing fine-tuning

We evaluate methods which have been designed to improve model robustness. In particular, we evaluate RXF (Aghajanyan et al., 2021) and FreeLB (Zhu et al., 2020), which apply adversarial training strategies to improve generalization. Previous work evaluated robustness by comparing only to a few models and do not run extensive evaluations in question answering. Our work conducts evaluations on a large number of distribution shifts.

2.2.5 In-context learning

In-context learning is an adaptation method proposed by Brown et al. (2020a) that does not require any gradient updates. This is particularly useful for very large language models, where fine-tuning is expensive. In-context learning refers to the process of conditioning a language model on one or more samples from a training set at inference time, allowing the model to perform a task without updating any parameters. For our experiments, we condition the model on triplets of context, question, and answer, as in Brown et al. (2020a).

2.2.6 Zero-shot inference

We evaluate models using prompting or zero-shot inference (Radford et al., 2019), where a model is conditioned only on the context and question of each test example. In other words, the model generates an answer without conditioning on training examples. Zero-shot models do not observe data from the reference distribution and have been shown to exhibit consistent performance across many distributions in computer vision (Radford et al., 2021).

2.3 Distribution shifts

We consider models which are trained on a reference distribution, which we also refer to as the in-distribution, with the exception of zero-shot models. In addition to measuring model performance on this reference distribution, we also evaluate model performance on other datasets where data distribution changes from the reference distribution. We refer to these other datasets as out-of-distribution, and we are interested in model behavior under distribution shift. Concretely, we want to measure how model performance changes when evaluated in- and out-of-distribution.

While there is extensive literature studying adversarial distribution shifts (Wu et al., 2021), our work focuses on *natural* distribution shifts (Taori et al., 2020), where the out-of-distribution datasets are not generated via synthetic perturbations to existing datasets.

In this work, we use the popular SQuAD (Rajpurkar et al., 2016) dataset as the reference (in-distribution) dataset. In addition, we evaluate model performance on 15 out-of-distribution datasets. We choose SQuAD as the reference distribution as it is one of the largest and the most well-studied QA datasets.

For our out-of-distribution test sets, we use the four datasets presented in the SquadShifts (Miller et al., 2020) in addition to datasets from the MRQA (Fisch et al., 2019) testbed. Details about each of these datasets can be found in Table 1.

2.4 Measuring robustness

We follow the technique for measuring model robustness that is outlined in Taori et al. (2020): a model is said to be robust if it exhibits consistent performance in- and out-of-distribution. This is advantageous compared to examining only out-of-distribution performance because it removes the confounder of in-distribution performance (as shown in (Taori et al., 2020; Miller et al., 2021b), models which achieve better performance in-distribution will often also perform better out-of-distribution).

As in Taori et al. (2020), the robustness measure we consider can be illustrated by looking at a scatter plot. For an illustrated example of this we refer to Figure 2, which displays the F1 score on the SQuAD development set on the x-axis and the F1 score averaged over the out-of-distribution datasets on the y-axis. Each point on the scatter plot is a different model. *Effective robustness* then describes *vertical* movement in this scatter plot towards the $y = x$ line. In particular, effective robustness measures performance out-of-distribution beyond the trend fit to fully fine-tuned models. This vertical movement is movement towards a model that has consistent performance in- and out-of-distribution (i.e., on aggregate fully fine-tuned models have $\sim$0 effective robustness). In Figure 2, which schematizes results that we will later observe with real data, models that are more robust sit above the baseline trend and exhibit robustness—the models shown in orange are more robust than the other models as they have better out-of-distribution performance given the same in-distribution performance.

3 Results

This section aims to answer our main research questions:

- How do models perform under distribution shift?
- Are some models more robust than others?
- Do adaptation methods impact robustness?

We answer these questions in Sections 3.1, 3.2 and 3.3, respectively.

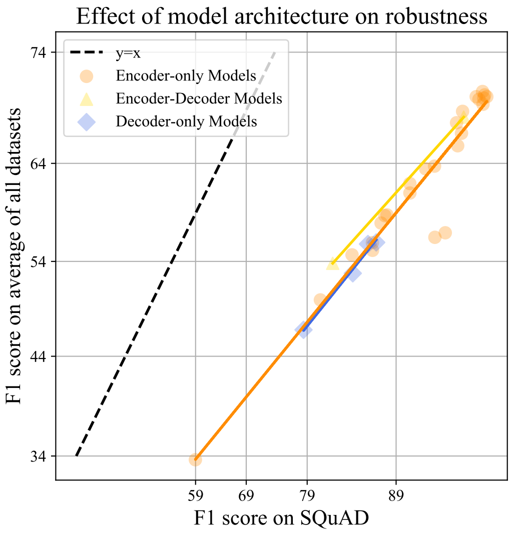


Figure 3: Encoder-only, encoder-decoder, and decoder-only models are equally as robust when fine-tuned by adding a span prediction head. We conclude that architecture does not determine distributional robustness.

3.1 Performance drops under distribution shift

As shown in Figure 1, we observe that model performance drops under distribution shift. This effect is more pronounced for the best models on SQuAD, which are fully fine-tuned. This indicates that, despite progress in question answering, there is still substantial room for progress in improving model robustness.

3.2 Role of model

Role of model architecture. In Figure 3 we compare the robustness of fine-tuned encoder-only, decoder-only, and encoder-decoder architectures. Our experiments indicate that architecture does not impact robustness. We observe that when different model families are adapted using a span prediction head, all models are equally robust. One limitation in our comparison is that the architectures we compare do not share the same pre-training corpus. However, larger corpora have been shown to improve robustness in computer vision (Radford et al., 2021). This is an area that could be investigated further in future work.

Role of model size. Previous work (Hendrycks et al., 2020) has claimed that model size does not affect the robustness of language models. In Figure 5 we plot the average *effective robustness* on all distribution shifts as a function of the number of

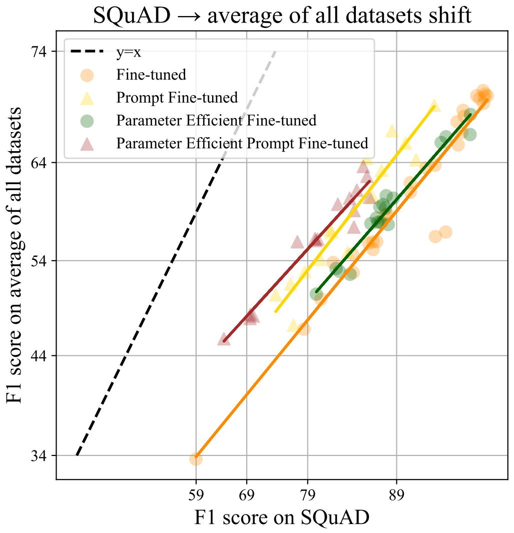


Figure 4: Parameter-efficient fine-tuning methods (highlighted in red and green) do not exhibit noticeable robustness improvements compared to other fine-tuned models.

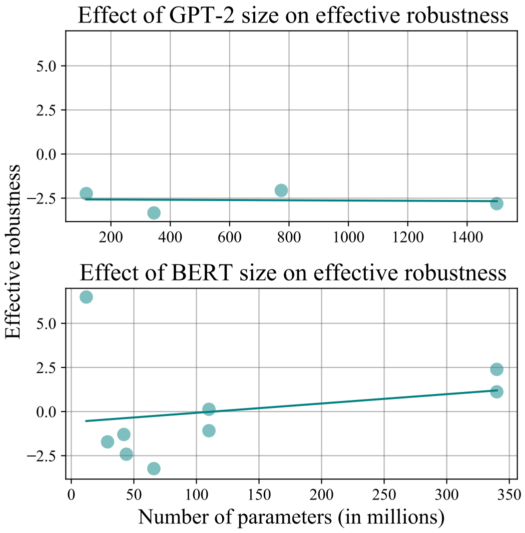


Figure 5: Average effective robustness of BERT and GPT-2 as a function of the number of parameters of models fine-tuned on SQuAD. Overall model size does not determine robustness.

model parameters for fine-tuned GPT-2 and BERT models to control for pre-training corpus and architecture. Overall, we observe that model size is not strongly correlated with robustness.

3.3 Role of the adaptation method

Zero-shot and in-context learning (ICL). We find that both zero-shot and in-context learning methods exhibit more robustness than methods that use gradient-based learning. As illustrated by Figure 1, the trend for zero-shot and in-context learning models is well above the trend of all other models. This entails that for the same in-distribution performance, we expect better out-of-distribution performance for in-context learning and zero-shot inference

Few-shot fine-tuning. In Figure 1, we observe that few-shot methods follow two separate robustness trends.

1. Few-shot fine-tuned models are on a trend similar to fully fine-tuned models.
2. Few-shot prompt fine-tuned models are more robust than all other models that use gradient based learning.

Notable outliers to the few-shot prompt fine-tuned model trend are the GPT-2 XL (Radford et al., 2019) and GPT-Neo 1.3B (Black et al., 2021) models. As shown in Figure 7, these models are more robust than other few-shot prompt fine-tuned models. This indicates that models with better zero-shot capabilities can generalize better when fine-tuned in the few-shot setting. For these few-shot fine-tuned GPT models we explore how the number of training shots impacts robustness. We find that as the number of training samples increases, the effective robustness of few-shot GPT models decreases as shown in Figure 8. In particular, increasing the number of shots from 16 to 1024 decreases effective robustness. This observation interpolates our previous results: a GPT model used in the zero-shot setting is robust while prompt fine-tuned GPT models are less robust. As observed by previous work (Radford et al., 2021; Andreassen et al., 2021; Wortsman et al., 2021), fine-tuning a model can reduce robustness and lead to a model which is overspecialized to the downstream task.

Full fine-tuning using span prediction. The fully fine-tuned models exhibit noticeably less robustness than other adaptation methods, however they also have the best performance on SQuAD. The best performing model on SQuAD has similar performance out-of-distribution to the best ICL model, despite performing more than 10 percentage points better in-distribution.

Fine-tuning using a prompt. We find that prompt fine-tuning methods are more robust in comparison to fine-tuned models. We observe that

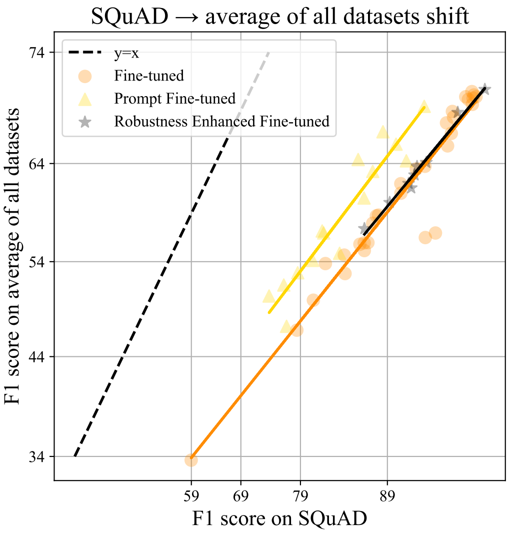


Figure 6: Methods designed to improve robustness (highlighted in black) do not exhibit noticeable robustness improvements on our testbed. This discrepancy may arise because of our focus on question answering, which previous work does not evaluate on.

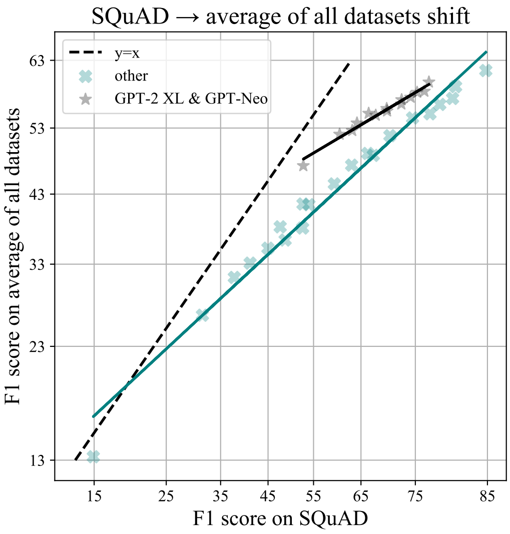


Figure 7: Few-shot prompt fine-tuned billion parameter GPT models (colored black) are more robust than smaller few-shot prompt fine-tuned models. Further investigation is required to determine if the increase in effective robustness is due to architecture or model size.

not using span prediction and instead fine-tuning existing model weights to generate the answer allows the model to maintain some of robustness from the zero-shot setting.

Parameter-efficient tuning. We examine the performance of parameter-efficient fine-tuning methods for different architectures and model sizes. Our results indicate that these methods are neither noticeably more robust or less robust than fine-tuning all parameters when using prompt based methods or span prediction, as shown in Figure 4.

Methods designed to enhance robustness. As illustrated by Figure 6 we find that RXF and FreeLB, which are designed to improve robustness, do not exhibit noticeable robustness improvements on the distribution shifts. We believe that one of the values of our large test bed is to comprehensively evaluate future robustness enhancing methods.

4 Discussion

This section discusses the aforementioned findings. In particular we discuss how the findings compare to analogous studies in vision, and how individual distribution shifts differ from aggregate trends.

4.1 How do the findings compare to robustness evaluations in vision?

We observe that the overall robustness trends of question answering models are qualitatively similar to trends identified in image classification (Taori et al., 2020; Radford et al., 2021; Miller et al., 2021a). In particular, previous work (Radford et al., 2021; Wortsman et al., 2021; Pham et al., 2021) has shown that zero-shot models are more robust than fine-tuned models, which is similar to the trend we observe. Moreover, additional robustness evaluations (Taori et al., 2020) have concluded that fully trained models models with different architectures, pre-training datasets, and robustness enhanced methods do not provide any robustness improvement when evaluated on multiple natural distribution shifts, which is also what we observe.

4.2 How do individual distribution shifts differ from aggregate trends?

While we have previously analyzed robustness trends averaged over all distribution shifts, we now examine trends on individual distribution shifts. For most datasets, we observe qualitatively similar trends as when averaging over all distribution shifts.

One exception is on the SquadShifts New-Wiki dataset, where we find that all models sit very

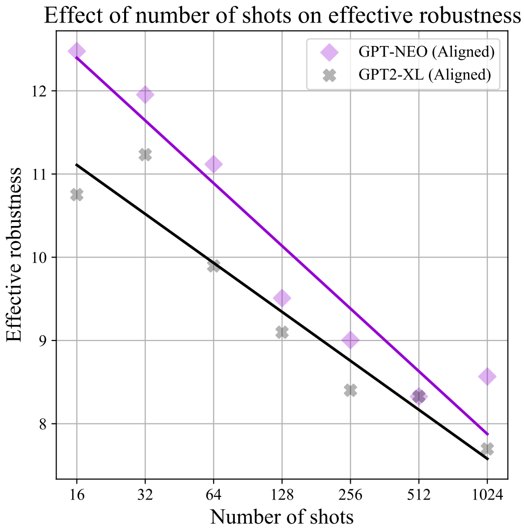


Figure 8: Average effective robustness for each GPT model as a function of the number of shots used for fine-tuning. As the number of shots increases the model becomes better in distribution but the average effective robustness decreases.

close to the $y = x$ line (Figure 9a). Since both SQuAD and SquadShifts New-Wiki are collected from Wikipedia, it is perhaps unsurprising that models adapted to SQuAD can generalize to other datasets from the same domain.

Moreover, we observe a piece-wise linear trend when comparing few-shot and fine-tuned models on DROP 9c. By fine-tuning on the entire training set, we improve in-distribution performance, which causes larger gains in DROP performance. Similar patterns of discontinuous improvement have been previously observed by Wei et al. (2022).

Additionally we find that on the SearchQA 9b dataset the trendlines are flatter than other distribution shifts for fine-tuned, prompt fine-tuned and few-shot fine-tuned models (i.e., increasing ID performance has a smaller impact on OOD performance). In addition, zero-shot and ICL models do not have additional robustness properties. The exception to this is GPT-J and OPT 175B, which continue to outperform other models. Moreover, few-shot prompt fine-tuned models perform better than other few-shot and fine-tuned models.

5 Related work

Understanding how models behave under conditions that differ from training has been the subject of much attention both in natural language processing (Ribeiro et al., 2020; Tu et al., 2020; Hendrycks et al., 2020; Gardner et al., 2020; Arora et al., 2021; Veitch et al., 2021; Goel et al., 2021; Miller et al., 2020, inter alia) and computer vision (Recht et al., 2019; Taori et al., 2020; Miller et al., 2021a; Koh et al., 2021, inter alia). As in Taori et al. (2020), we distinguish between synthetic and natural distribution shifts. The former includes any artificial perturbations to inputs, including adversarial attacks (Szegedy et al., 2013; Carlini and Wagner, 2017; Jia and Liang, 2017; Biggio and Roli, 2018; Wang et al., 2019b; Wallace et al., 2019b,a; Tramer et al., 2020; Liu et al., 2021; Wu et al., 2021; Chang et al., 2021, inter alia). In contrast, the later relates to naturally occurring data, without synthetic or adversarial perturbations. Our work focuses on natural distribution shifts.

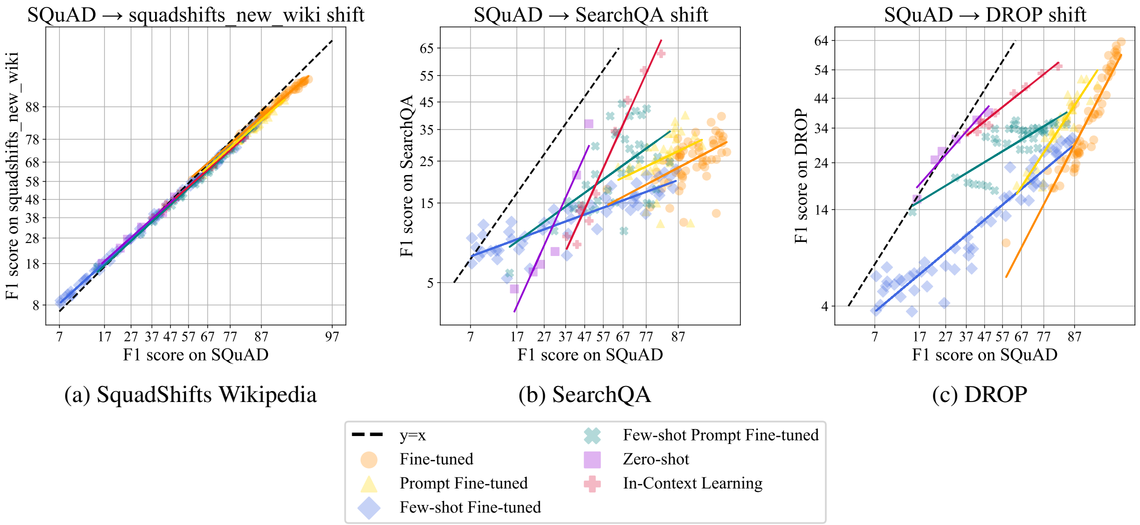
Most similar to our work is that of Yogatama et al. (2019); Talmor and Berant (2019); Sen and Saffari (2020); Fisch et al. (2019) and Miller et al. (2020), who examine the performance of models on multiple question answering datasets. Our work provides a more comprehensive modeling survey, evaluating a broader set of models, adaptation strategies and datasets. In contrast to previous work, we evaluate zero-shot inference, in-context learning, few-shot fine-tuning, and parameter-efficient adaptation methods, which have only recently been popularized.

Finally, a variety of methods for improving robustness have been explored by previous work (Jiang et al., 2019; Zhu et al., 2020; Aghajanyan et al., 2021; Veitch et al., 2021; Wortsman et al., 2021, inter alia). Instead of proposing methods to build more robust models, our goal is to empirically examine the landscape of robustness. As part of this goal, we evaluate robustness enhancing methods, in addition to other adaptation strategies.

Concurrent work by Liu et al. (2022) examines the robustness of few-shot fine-tuned models. They find that these models yield no additional robustness which matches the findings from our evaluation.

6 Conclusion

We conduct an extensive evaluation of the robustness of different model and adaptation methods on 15 distribution shifts in question answering. Our in-depth analysis suggests several concrete directions for future work: improving the in-distribution performance of ICL methods and understanding why different few-shot fine-tuning methods yield



(a) SquadShifts Wikipedia

(b) SearchQA

(c) DROP

Figure 9: Instead of averaging over all 15 datasets, we now show logit-scaled plots examining the three distribution shifts individually. **(left)** The SquadShifts Wiki dataset is derived from the same data source (Wikipedia) as SQuAD. As a result, models lie closer to the $y = x$ diagonal than on other distribution shifts. **(middle)** Progress on SQuAD is a weaker indicator for progress on SearchQA for fully fine-tuned models and few-shot fine-tuned models. We find that zero-shot and ICL models are less robust than fine-tuned and few-shot models with the exception of larger language models. **(right)** On the SQuAD→DROP distribution shift, we observe that progress beyond 70 F1 on SQuAD yields quick progress on DROP for fine-tuned models.

varied robustness.

7 Acknowledgements

This work is in part supported by the NSF AI Institute for Foundations of Machine Learning (IFML), Open Philanthropy, Google, and the Allen Institute for AI.

8 Limitations

Experimenting with different in-distribution datasets. We choose SQuAD as a representative in-distribution dataset since it is one of the largest and most popular QA datasets. One limitation of SQuAD is that the training set is mainly collected from Wikipedia articles which may not be optimal for building a QA model that generalizes to many domains. Future work could explore the robustness of models trained on datasets from other domains for increased coverage.

Specialized modeling methods. Our work does not evaluate models with task or data specific components. As an example Andor et al. (2019) improved performance on DROP (Dua et al., 2019) by using arithmetic programs to improve a model's mathematical reasoning. Evaluating the robustness of methods like these are an exciting area for future investigations.

Few-shot GPT evaluations. Our results indicate that large GPT models fine-tuned on a smaller number of samples are more robust to distribution shifts compared to other few-shot fine-tuned models that use a prompt or span prediction. However, GPT-2 XL and GPT-Neo, which both have more than one billion parameters, are larger than all few-shot models we evaluate. Future work could examine the impact of architecture on this trend by evaluating other models with more than a billion parameters like T5.

Multiple fine-tuning runs. For fine-tuned models we include a single data-point for each model. However, previous work (Phang et al., 2018; Dodge et al., 2020) has shown that different data ordering and weight initialization can lead to large variance in model performance. In Figure 11 we evaluate the robustness of RoBERTa Large models fine-tuned with different data ordering and initialization for the span prediction head (Devlin et al., 2019). We find that on average the robustness of these models does not differ substantially. Further investigation into the effect of random seeds on robustness would improve our understanding of the robustness of individual data points.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2021. Better fine-tuning by reducing representational collapse. In *International Conference on Learning Representations*.

Daniel Andor, Luheng He, Kenton Lee, and Emily Pitler. 2019. Giving BERT a calculator: Finding operations and arguments with reading comprehension. 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 5947–5952, Hong Kong, China. Association for Computational Linguistics.

Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. 2021. The evolution of out-of-distribution robustness throughout fine-tuning. *arXiv preprint arXiv:2106.15831*.

Udit Arora, William Huang, and He He. 2021. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee.

Rakesh Chada and Pradeep Natarajan. 2021. FewshotQA: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6081–6090, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kai-Wei Chang, He He, Robin Jia, and Sameer Singh. 2021. Robustness and adversarial examples in natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 22–26.

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*.

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

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah A. Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*, abs/2002.06305.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

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

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 2nd Machine Reading* *for Reading Comprehension (MRQA) Workshop at EMNLP*.

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.

Karan Goel, Nazneen Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328*.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2019. Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. *arXiv preprint arXiv:1911.03437*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

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.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

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

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

Nelson F. Liu, Ananya Kumar, Percy Liang, and Robin Jia. 2022. Are sample-efficient nlp models more robust? *ArXiv*, abs/2210.06456.

Nelson F Liu, Tony Lee, Robin Jia, and Percy Liang. 2021. Can small and synthetic benchmarks drive modeling innovation? a retrospective study of question answering modeling approaches. *arXiv preprint arXiv:2102.01065*.

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*.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2021. Time waits for no one! analysis and challenges of temporal misalignment. *arXiv preprint arXiv:2111.07408*.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *ICML*.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021a. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, pages 7721–7735. PMLR.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021b. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7721–7735. PMLR.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. 2021. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

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.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021a. Few-shot question answering by pretraining span selection. In *ACL*.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021b. Few-shot question answering by pretraining span selection. 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 3066–3079, Online. Association for Computational Linguistics.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR.

Marco Tulio Ribeiro and Scott Lundberg. 2022. Adaptive testing and debugging of NLP models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3253–3267, Dublin, Ireland. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. Duorc: Towards complex language understanding with paraphrased reading comprehension. In *ACL*.

Priyanka Sen and Amir Saffari. 2020. What do models learn from question answering datasets? *arXiv preprint arXiv:2004.03490*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Alon Talmor and Jonathan Berant. 2019. Multiqa: An empirical investigation of generalization and transfer in reading comprehension. *arXiv preprint arXiv:1905.13453*.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. 2020. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. 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.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael Alvers, Dirk Weißenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. Advances in Neural Information Processing Systems.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. 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 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*, 7:387–401.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. 2019b. Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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.

Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*.

Winston Wu, Dustin Arendt, and Svitlana Volkova. 2021. Evaluating neural model robustness for machine comprehension. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2470–2481, Online. Association for Computational Linguistics.

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.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

As at most other universities, Notre Dame s students run a number of news media outlets. The nine student-run outlets include three newspapers, both a radio and television station, and several magazines and ournals.
Question: *How many student news papers are found at Notre Dame*
Answer: *three*

Figure 10: A sample from SQuAD with the input formatting used for fine-tuning decoder-only models, in-context learning, and zero-shot inference.

A Appendix

A.1 Training Details

In addition to sharing hyperparameters for each model, we plan to share all model weights on the HuggingFace Hub (Wolf et al., 2020) such that the community can continue to evaluate the models in our testbed.

A.1.1 Span prediction fine-tuning

We fine-tune models by adding a span prediction head and fine-tuning for 2 epochs using a learning rate of 3e-5 and a linear learning rate decay.

A.1.2 Prompt fine-tuning

Encoder-Decoder Models
We fine-tune encoder-decoder models on both question->answer generation (mask filling) and answer generation tasks from Chada and Natarajan (2021). For the question->answer generation task we fine-tune the models for 2 epochs, use a linear learning rate decay, and search for the best learning rate from 1e-4, 5e-5, and 3e-5 based on performance on the validation set. For the answer generation task we fine-tune for 2 epochs with a learning rate of 3e-5 and linear learning rate decay.
Decoder-Only Models
We fine-tune decoder only models using a language modeling head. Specifically we format samples as shown in Figure 10 and only calculate loss on the answer tokens. We search for the best learning rate among 5e-5 and 5e-6 and use a linear learning rate decay. In addition we search for the best weight decay value between 0.01 and 0.1. We fine-tune these models for 5 epochs and pick the model with the best validation set F1 score.

A.1.3 Parameter efficient fine-tuning

Fine-tuned Models
As suggested in the Adapter-Transformers library (Pfeiffer et al., 2020) we use a learning rate of 1e-4 with linear learning rate decay and fine-tune for 15 epochs for all parameter efficient fine-tuning methods picking the model with the best F1 score on the validation set.
Prompt Fine-tuned Models
We fine-tune using a learning rate of 1e-4 with linear learning rate decay and 10 epochs for all parameter efficient fine-tuning methods picking the model with the best F1 score on the validation set.

A.1.4 Few-shot Fine-tuning

We fine-tune models on 2^4 to 2^{10} samples from SQuAD (doubling the size as we increase the number of shots). We repeat each experiment three times using randomly picked samples to remove outliers that result from fine-tuning on specific examples.
Fine-tuned Models
We use the same fine-tuning setup as Ram et al. (2021b). Specifically we fine-tune for 10 epochs or 200 steps (picking which ever is largest). We use a learning rate of 3e-5 with a linear learning rate decay and 0.1 warm-up ratio.
Prompt Fine-tuned Models
For autoregressive models we fine-tune for 10 epochs with a learning rate of 1e-5 and linear learning rate decay. In addition we use a weight decay of 0.1 as we find in our experiments fine-tuning decoder-only models that this is an ideal value for the models we evaluate. For T5 and BART we use the same evaluation setup as Chada and Natarajan (2021) for both masked span prediction and answer generation methods.

A.1.5 Robustness enhanced methods

RXF
We adapt the official implementation[1] for RXF (using the R3F variant) to fine-tune encoder only question answering models. We use the same fine-tuning hyper-parameters as the fully fine-tuned encoder only models A.1.1 but use polynomial learning rate decay, weight decay value of 0.01, and warm-up ratio of 0.06. For R3F specific parameters we use λ=1.0, δ=1e-5, and Normal noise type.
FreeLB
We use the official implementation[2] for FreeLB to fine-tune encoder only question answering models. We use a learning rate of 5e-6 and fine-tune for 2 epochs with linear learning rate decay. For FreeLB

[1] https://github.com/facebookresearch/fairseq/tree/main/examples/rxf
[2] https://github.com/zhuchen03/FreeLB

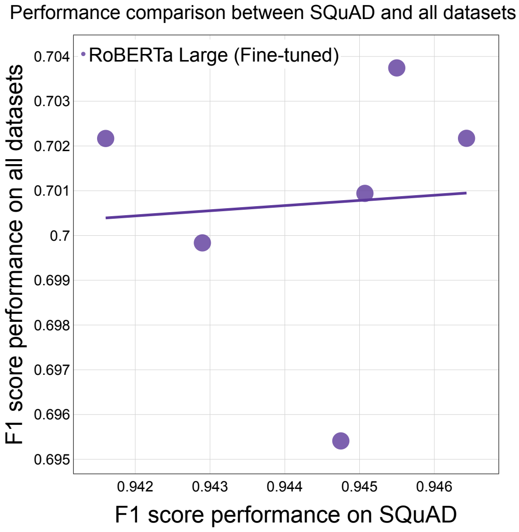


Figure 11: The average effective robustness for six fine-tuning runs of RoBERTa Large shows that the robustness differences between fine-tuning runs are negligible.

specific parameters we set $m = 2$, α=1e-1, and ϵ=6e-1.

A.1.6 Zero-shot inference

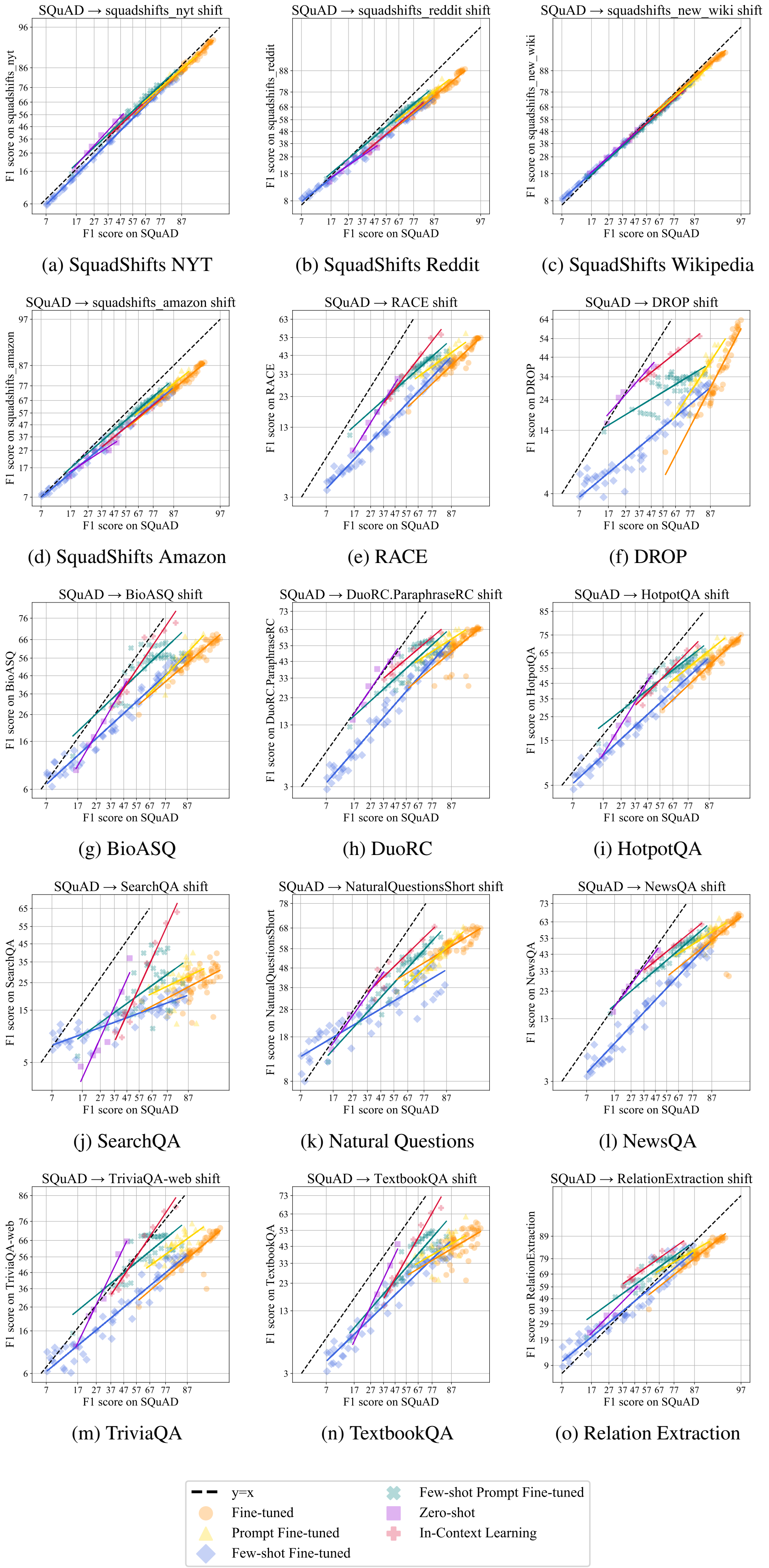
For zero-shot evaluations we pre-process each sample into the format in Figure 10 omitting the answer from the prompt. We generate the answer using beam decoding with five beams for models smaller than 2 billion parameters and use greedy decoding for the rest of the models. We use a maximum generation length of 20 tokens and use the end of sequence token to terminate generation.

A.1.7 In-context learning

For in-context learning evaluations we condition a language model on one or four random samples examples from the SQuAD training set. Figure 10 illustrates the format of each sample. When we are conditioning on multiple samples we separate each formatted sample with a newline delimiter. For each training sample we condition on, we truncate the context of the sample to 100 tokens. Furthermore, we truncate the context of input sample (sample we are running inference on) to 200 tokens. For each model and number of shots we repeat each experiment three times using randomly samples training shots. The exception to this is the OPT 175 billion parameter model which we evaluate only once in the one and four shot settings. We use the same generation setup as zero-shot inference A.1.6.

A.2 Additional Plots

In this section we include Figure 11 which shows the effect of different fine-tuning runs on effective robustness. We find that even when we fine-tune six different RoBERTa Large models by varying data ordering and weight initialization for the span prediction head the average effective robustness on all distribution shifts is stable.



(a) SquadShifts NYT

(b) SquadShifts Reddit

(c) SquadShifts Wikipedia

(d) SquadShifts Amazon

(e) RACE

(f) DROP

(g) BioASQ

(h) DuoRC

(i) HotpotQA

(j) SearchQA

(k) Natural Questions

(l) NewsQA

(m) TriviaQA

(n) TextbookQA

(o) Relation Extraction

Figure 12: Instead of averaging over all 15 datasets, we show logit-scaled plots examining all 15 distribution shifts individually.