Can Out-of-Distribution Evaluations Uncover Reliance on Shortcuts? A Case Study in Question Answering

Michal Štefánik[♡♣*] **Timothee Mickus**[♡]
Marek Kadlčík[♣] **Michal Spiegel**[♣◇] **Josef Kuchař**[♣]

[♡]Language Technology, University of Helsinki
[♣]TransformersClub @ Faculty of Informatics, Masaryk University
[◇]Kempelen Institute of Intelligent Technologies

Abstract

A majority of recent work in AI assesses models' generalization capabilities through the lens of performance on out-of-distribution (OOD) datasets. Despite their practicality, such evaluations build upon a strong assumption: that OOD evaluations can capture and reflect upon possible *failures* in a real-world deployment.

In this work, we challenge this assumption and confront the results obtained from OOD evaluations with a set of specific failure modes documented in existing question-answering (QA) models, referred to as a reliance on spurious features or prediction shortcuts.

We find that different datasets used for OOD evaluations in QA provide an estimate of models' robustness to shortcuts that have a *vastly* different quality, some largely underperforming even a simple, in-distribution evaluation. We partially attribute this to the observation that spurious shortcuts are *shared* across ID+OOD datasets, but also find cases where a dataset's quality for training and evaluation is largely disconnected. Our work underlines limitations of commonly-used OOD-based evaluations of generalization, and provides methodology and recommendations for evaluating generalization within and beyond QA more robustly.

Models Rank

Out-of-Distribution: SQuAD-valid … Avg: OOD; Reliance on shortcuts: Q-words dist … Avg: shortcuts

	SQuAD-valid	TriviaQA	AdversarialQA	NewsQA	SearchQA	NaturalQuestions	**Avg: OOD**	Q-words dist	Entity match	Keywds match	Shared words	Answer length	**Avg: shortcuts**
ahotrod/ELECTRA-L	1	2	3	3	1	5	1	3	3	1	4	2	1
RoBERTa-L (ours)	3	4	4	2	7	7	5	4	2	2	2	5	2
T5-L (ours)	5	3	2	1	5	1	2	1	4	4	3	9	3
deepset/DeBERTa3-L	1	1	1	4	7	8	3	2	5	3	8	3	3
deepset/XLM-R-L	8	5	5	5	2	10	6	8	7	10	1	1	5
deepset/RoBERTa-B	6	7	7	8	4	4	7	6	1	8	7	6	6
deepset/RoBERTa-Dist	4	6	6	5	3	2	4	5	6	7	6	8	7
RoBERTa-B (ours)	7	8	8	7	10	3	8	7	9	5	5	7	8
ELECTRA-B (ours)	9	9	9	9	9	6	9	9	8	9	9	4	9
BERT-B (ours)	10	10	10	10	6	9	10	10	10	6	10	10	10

Figure 1: **How to pick the most robust model?** Ranking of popular QA models by two facets of generalization: Out-of-Distribution evaluations (left) and Reliance on prediction shortcuts/spurious features (right); models are ordered by the average ranking across all spurious features (last column).

1 Introduction

Improving the generalization of language models (LMs), i.e., their capability to perform well beyond patterns covered by their limited training data (Chollet, 2019; Guo et al., 2023), presents one of the most important challenges in modern NLP, with direct implications to their practical applicability in a wide variety of tasks. The most common approach towards evaluating LMs' ability to generalize is to assess their performance on so-called out-of-distribution (OOD) datasets: datasets of the same task(s) but of different origins — an approach made practical by the large variety of pre-existing datasets for an extensive array of NLP tasks. We expect a model that generalizes well to achieve high scores on OOD data. However, as we hold limited knowledge of the properties of datasets, it is difficult to ensure that OOD datasets can comprehensively capture real-world failures to generalize observed in LMs' practical applications.

On the other hand, a growing body of work has been documenting and classifying the systematic generalization failures of LMs as due to *prediction bias* (Utama et al., 2020), *prediction shortcuts* (Mikula et al., 2024), or a reliance on *spurious features* (Zhou et al., 2021). These failures are characterized by a model's over-reliance on a feature that can explain the training data well but is

[*]From October 2025 at National Institute of Informatics Research and Development Center for Large Language Models, Tokyo, Japan

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22628–22635
November 4-9, 2025 ©2025 Association for Computational Linguistics

not representative of the task overall.

In this case study, we confront these two orthogonal views on generalization — OOD performance and non-reliance on shortcuts — by using them to establish two sets of independent rankings for a selection of the most popular QA models. We focus on extractive QA where previous work provides the most extensive documentation of prediction shortcuts, allowing for a robust assessment.

We find that many OOD datasets can accurately portray models' robustness to shortcuts, but an uninformed selection of OOD datasets can also deliver a ranking that is *not at all* correlated with a reliance on shortcuts, underperforming even traditional in-domain evaluations. Finally, by assessing reliance on shortcuts for models trained on datasets used as OOD, we show that different QA datasets exhibit the *same types* of shortcuts. Nevertheless, we find that a dataset's usefulness in uncovering shortcuts does *not* entail that it can be used to train more robust models.

2 Background

Evaluating generalization in QA Evaluations on out-of-distribution (OOD) datasets (Wang et al., 2022) are decisively the most popular method for evaluating generalization of language models — to the point that performance on OOD is often even referred *interchangeably* with the term of generalization (Yang et al., 2023a). Within QA, among many others, Awadalla et al. (2022) train models on SQuAD (Rajpurkar et al., 2016) as in-distribution (ID) and evaluate generalization on TriviaQA (Joshi et al., 2017), NewsQA (Trischler et al., 2017), SearchQA (Dunn et al., 2017) or NaturalQuestions (Kwiatkowski et al., 2019) as OOD. Clark et al. (2019) evaluates even shortcut-eliminating method by training on SQuAD and evaluating on AdversarialQA (Jia and Liang, 2017) as OOD. Yogatama et al. (2019) train on SQuAD as ID and evaluate on TriviaQA as OOD. Despite its known blindspots, SQuAD still remains the default training dataset for a majority of the most popular QA models on HuggingFace. With a primary objective of improving generalizations, these works *assume* that OOD evaluations can also uncover reliance on non-representative, spurious features.[1] In this work, we question this assumption and find cases where OOD evaluations are largely *independent* from a reliance on non-representative, spurious features.

Prediction shortcuts A complementary yet less prevalent approach to assessing models' generalization aims to exploit functional failures identified in previous models. One approach towards this goal consists in identifying models' *prediction shortcuts*, i.e. a reliance on *spurious features* that are not representative for the learned task in general. Such shortcuts were previously identified in NLI (Nie et al., 2020), in-context learning (Wei et al., 2023) or question answering (Mikula et al., 2024). While these shortcuts are difficult to identify, with their knowledge, we can test the model specifically for the reliance on each of the shortcuts. This can be done by *constructing* synthetic data (Clark et al., 2019) or *subsetting* existing data (Mikula et al., 2024) into subset(s) where we make sure that a specific shortcut is *not* applicable.

In NLI, where the impact of specific prediction shortcuts is widely studied, a common practice is to evaluate on datasets specifically constructed to exploit reliance on shortcuts (McCoy et al., 2019). However, in open-ended tasks including QA or multitask benchmarks such as MMLU (Hendrycks et al., 2021), the applicability of shortcuts across different datasets used for OOD evaluations remains under-studied.

3 Methodology

Our goal is to uncover if, and to what extent can commonly-performed OOD evaluations capture models' failures attributed to models' reliance on prediction shortcuts. We approach this by *reproducing* several OOD evaluations used in previous work and *comparing* the results of these evaluations with models' measured sensitivity to prediction shortcuts representing previously reported failures modes.

Models For our assessments, we aim to pick a set of existing QA models that are the most widely-used in practice. Towards this goal and with a minor preference for diversity, we pick five among the thirty most-downloaded models on HuggingFace[2] with the added criterion that they must have been trained on SQuAD only, so as to ensure a fair comparison. Additionally, to further enhance diversity,

[1] While we focus on QA, we can easily find such an assumption in other tasks, including NLI (Du et al., 2021; Korakakis and Vlachos, 2023) or classification (Yang et al., 2023b).

[2] `https://huggingface.co/models?pipeline_tag=question-answering&sort=downloads`; accessed 21/04/25. The selected models total over 2.4 million monthly downloads.

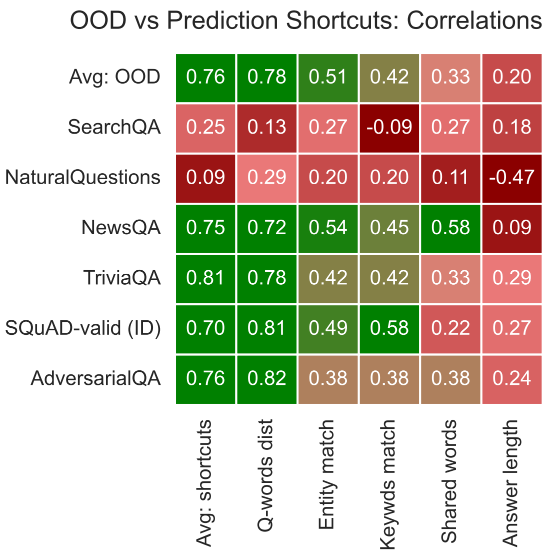

Figure 2: **Kendall τ Correlation** (a ratio of changed *orderings*) between the rankings provided by two different facets of generalization: (**rows**) performance on OOD datasets; and (**columns**) reliance on prediction shortcuts. First row and column contain an average rankings across all evaluated datasets and shortcuts.

we complement this selection by training a set of our own models covering the most popular model families among QA, such that all other training parameters (e.g, early stopping, batch size, learning rate) are fixed across different model families.

ID & OOD Datasets Consistent with previous work (§2) and the most popular QA models, we use SQuAD (Rajpurkar et al., 2016) as our ID dataset unless otherwise noted. To evaluate robustness, we pick five different OOD datasets, *all* used in prior studies. We then use the performances on each of these datasets, as well as their average, to establish a first set of rankings for our models.

Prediction shortcuts We create a second set of model rankings by evaluating their reliance on five shortcuts identified in previous work on QA and attested across all models in our study (detailed in §A): shared words (Shinoda et al., 2021), question-words distance (Jia and Liang, 2017), keywords match (Clark et al., 2019), answer length (Bartolo et al., 2020) and entity match (Mikula et al., 2024). To quantify the reliance of a given model on each of these shortcuts, we follow the methodology of Mikula et al. (2024). In practice, we (1) split the in-distribution validation dataset into two segments based on whether the example (i) can or (ii) can *not* be solved by the shortcut; (2) compute the accuracy on both segments; and (3) calculate the relative drop in accuracy between the segments.[3]

Metrics We assess the agreement between the OOD evaluations and models' reliance on shortcuts by visualizing the ranking obtained by each of these features, and their mutual correlation. All our reported results employ the exact-match metric.[4]

4 Results

Figure 1 displays model rankings as derived from the two approaches for evaluating generalization: (left) OOD evaluations relied upon in previous work (§2) and (right) reliance on a set of prediction shortcuts, i.e. a relative drop in models' accuracy when shortcuts are not applicable. The last column in each group presents an *average* within the group. Models marked as (*ours*) are newly-trained and mutually comparable while other models present the most popular QA models from HuggingFace.

Lower-position rankings are relatively consistent across both OOD and shortcuts. However, discrepancies become more pervasive at the *top* positions, instructive for picking the *most* robust model. Here, a majority (3/5) of OOD evaluations do *not* agree on the selection of the most robust model with the average ranking by shortcuts. Consequentially, picking the 'most robust' model based on some OOD datasets (NewsQA and NaturalQuestions) may yield the model with a 23% larger average dependence on shortcuts. Noticeably, OOD evaluations on these datasets rank the highest those models that rely on shortcuts *more* than those considered the best by a standard ID evaluation.

In Figure 2, we compare the different rankings via Kendall τ pairwise correlations, a metric proportional to the number of pairwise swaps needed to *transform* one ranking into another. Results reveal that there are two vastly different OOD datasets: NaturalQuestions and SearchQA. Rankings according to these datasets have *minimal* correlation ($\tau < 0.4$) with *each* of the shortcuts, but, as we find, also other with datasets. Both datasets correlate with the averaged reliance on shortcuts substantially *worse* than the traditional ID evaluations. On the other hand, some OOD evaluations correlate with average shortcuts' ranking much better than ID or even averaged OOD performance —

[3]We ensure the statistical significance of our shortcuts through bootstrapped confidence intervals.

[4]We confirmed on a subset of our evaluations that compared to using F-score, the exact-match metric does not have any effect on the resulting ranking in a majority of evaluations. See §B.1 for absolute values for OOD evaluation and shortcut.

	TriviaQA	Adv'QA	NewsQA	SQuAD	SearchQA	NQ's
Eval (↑)	0.87	0.78	0.75	0.70	0.25	0.09
Train (↓)	15.28	2.79	3.09	3.17	16.21	3.22

Table 1: **Quality of datasets** for (top) **evaluation**, i.e. dataset's ability to uncover shortcuts in evaluation (reported in Fig. 2; ↑ higher is better), and (bottom) for **training**, i.e. a reliance on shortcuts for models trained on the given dataset (↓ smaller is better).

which suggests that ranking among a large set of OOD evaluations, assumed as more robust in some of previous works (Awadalla et al., 2022) need *not* provide a better assessment of robustness than an informed selection of a single OOD dataset.

4.1 Analyses

The large discrepancy among different evaluation datasets in their ability to uncover shortcuts raises the question of whether different datasets can indeed exhibit the *same* prediction shortcuts. If so, models' reliance on shortcuts would not only remain hidden from OOD evaluations but could even *improve* their OOD results. To answer this, we train new QA models on each of our OOD datasets and assess their reliance on each of our shortcuts. Except for the training data, our methodology remains identical to the training of ID models (§3). We limit our analyses to the model least reliant on shortcuts, viz. RoBERTa-Large.

In Table 1, we report two metrics for each of our datasets: (top row) the correlation with reliance on shortcuts when *evaluating* with the dataset (identical to Fig. 2), and (bottom row) the average reliance on shortcuts (a relative drop in accuracy) when *training* on this dataset. The datasets (x-axis) are ordered based on their correlation with ranking based on the reliance on shortcuts (Fig. 2).

Can different OOD datasets exhibit the *same* prediction shortcuts? Table 1 shows that except for AdversarialQA and NewsQA, models trained on OOD datasets rely on shortcuts of SQuAD similarly or even *more* than a SQuAD-trained model. In the case of TriviaQA and SearchQA, the drop in accuracy caused by the unavailability of shortcuts is around *five times larger* than that of the SQuAD model. However, detailed results (Appx. B.1) reveal that even the less-reliant AdversarialQA and NewsQA-trained models exhibit a significant reliance on shortcuts in the case of three and four out of seven inspected shortcuts. Together, these evaluations provide evidence that datasets used for OOD evaluations in previous work exhibit the *same* types of prediction shortcuts as the training data.

Are bad *training* datasets also bad *evaluation* datasets? We showed that different datasets can provide vastly different quality in both training robust models and uncovering reliance on shortcuts. Consequentially, we may assume that there is a *proportional* relationship between the dataset's ability to train a robust model and to uncover the reliance on shortcuts. However, relating these two facets in Table 1 (*Train* vs *Eval* row), we can see *no* clear relation between the dataset's quality for training and evaluation; For instance, while TriviaQA is the best proxy for evaluating models' reliance on shortcuts, using it as a training dataset delivers a model almost five times more reliant on shortcuts than less reliable SQuAD or NaturalQuestions (NQ's).

These results point to the presence of *other* covariates that determine datasets' quality independently for training and evaluation. We investigate several potential features, including dataset size, context size, and sample format. While we do not identify a robust discriminant in training, in evaluation, we find that the least robust evaluations are delivered by datasets with more specific formats containing delimiters of different context sections (SearchQA) or in-context references (NQ's). We hypothesize that while the context format may not necessarily harm the robustness in training, it may strongly bias the evaluation towards dominantly assessing robustness to the dataset-specific artifacts.

5 Conclusions

This paper investigates a discrepancy between OOD evaluations used in previous work in QA as a proxy for generalization, and previously documented failures to generalize identified as uncover reliances on prediction shortcuts. By ranking a set of popular QA models according to these two facets, we find that datasets previously used for OOD evaluations vastly differ in their capacity to uncover shortcuts.

We find that prediction shortcuts are to a large extent *shared* across datasets used for both training and evaluations of generalization. However, the quality of datasets in uncovering shortcuts is *not* proportional to their capacity to train more robust models, possibly due to dataset-specific features with a different impact on training and evaluation.

We hope our results will inspire future work towards a more systematic selection of OOD datasets and provide concrete recommendations for OOD datasets selection in QA. Crucially, our findings may also motivate future work in generalization within and beyond QA to restrain from overgeneralized conclusions based on demonstratedly limited OOD benchmarks.

Acknowledgements

This work is supported by the Research Council of Finland through project No. 353164 "Green NLP – controlling the carbon footprint in sustainable language technology". This project has received funding from the European Union's Horizon Europe research and innovation programme under Grant agreement No. 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (grant number 10052546). The contents of this publication are the sole responsibility of its authors and do not necessarily reflect the opinion of the European Union.

Limitations

We identify a primarily limitation of our work in a limited scope of prediction shortcuts that we survey. This limitation is conditioned by the scope of (seven) types of shortcuts identified in previous work, among which we identify only five to be significant for most of our evaluated models. This also restrains us from expanding our case study into other candidate tasks; in NLI, we identify only three known prediction shortcuts (McCoy et al., 2019), while in in-context learning, we identified in previous work only a single prediction shortcut to assess (*reliance on label's semantics* uncovered in Wei et al. (2023)).

We further acknowledge that the database of known prediction shortcuts presents only a small subset of functional failures, where the failures of other categories would certainly be also desirable to capture in generalization evaluations. This limitation invites future work to assess models for other notorious functional deficiencies as our knowledge of models' functioning will grow, in a methodology similar to ours.

References

Anas Awadalla, Mitchell Wortsman, Gabriel Ilharco, Sewon Min, Ian Magnusson, Hannaneh Hajishirzi, and Ludwig Schmidt. 2022. Exploring the landscape of distributional robustness for question answering models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5971–5987, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Max Bartolo, A Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

François Chollet. 2019. On the measure of intelligence. *ArXiv*, abs/1911.01547.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. 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 4069–4082, Hong Kong, China. ACL.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of NLU models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 915–929, Online. Association for Computational Linguistics.

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

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. 2023. Causal de finetti: On the identification of invariant causal structure in exchangeable data. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Michalis Korakakis and Andreas Vlachos. 2023. Improving the robustness of NLI models with minimax training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14339, Toronto, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proc. of the 57th Annual Meeting of the ACL*, pages 3428–3448, Florence, Italy. ACL.

Lukáš Mikula, Michal Štefánik, Marek Petrovič, and Petr Sojka. 2024. Think Twice: Measuring the Efficiency of Eliminating Prediction Shortcuts of Question Answering Models. In *Proceedings of the 18th Conference of the European Chapter of the ACL (Volume 1: Long Papers)*, pages 2179–2193, St. Julian's, Malta. ACL.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proc. of the 58th Annual Meeting of the ACL*, pages 4885–4901. ACL.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, USA. ACL.

Kazutoshi Shinoda, Saku Sugawara, and Akiko Aizawa. 2021. Can question generation debias question answering models? a case study on question-context lexical overlap. *arXiv preprint arXiv:2109.11256*.

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.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Mind the Trade-off: Debiasing NLU Models without Degrading the In-distribution Performance. In *Proc. of the 58th Annual Meeting of the ACL*, pages 8717–8729. ACL.

Xuezhi Wang, Haohan Wang, and Diyi Yang. 2022. Measure and Improve Robustness in NLP Models: A Survey. In *Proceedings of the 2022 Conference of the North American Chapter of the ACL: Human Language Technologies*, pages 4569–4586, Seattle, USA. ACL.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *Preprint*, arXiv:2303.03846.

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 *Proc. of the 2020 Conf. EMNLP: System Demonstrations*, pages 38–45. ACL.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023a. Out-of-distribution generalization in natural language processing: Past, present, and future. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Linyi Yang, Yaoxian Song, Xuan Ren, Chenyang Lyu, Yidong Wang, Jingming Zhuo, Lingqiao Liu, Jindong Wang, Jennifer Foster, and Yue Zhang. 2023b. Out-of-distribution generalization in natural language processing: Past, present, and future. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4533–4559, Singapore. 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*.

Chunting Zhou, Xuezhe Ma, Paul Michel, and Graham Neubig. 2021. Examining and Combating Spurious Features under Distribution Shift. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12857–12867. PMLR.

A Definitions of shortcuts

We include here a brief overview of the shortcut we study and refer the reader to the original works for relevant details.

- **Shared words** (Shinoda et al., 2021): Models' reliance on the assumption that correct answer must contain some words from the question.
- **Question-words distance** (Jia and Liang, 2017): Models' reliance on close proximity of the answer to some of the words contained in the question.

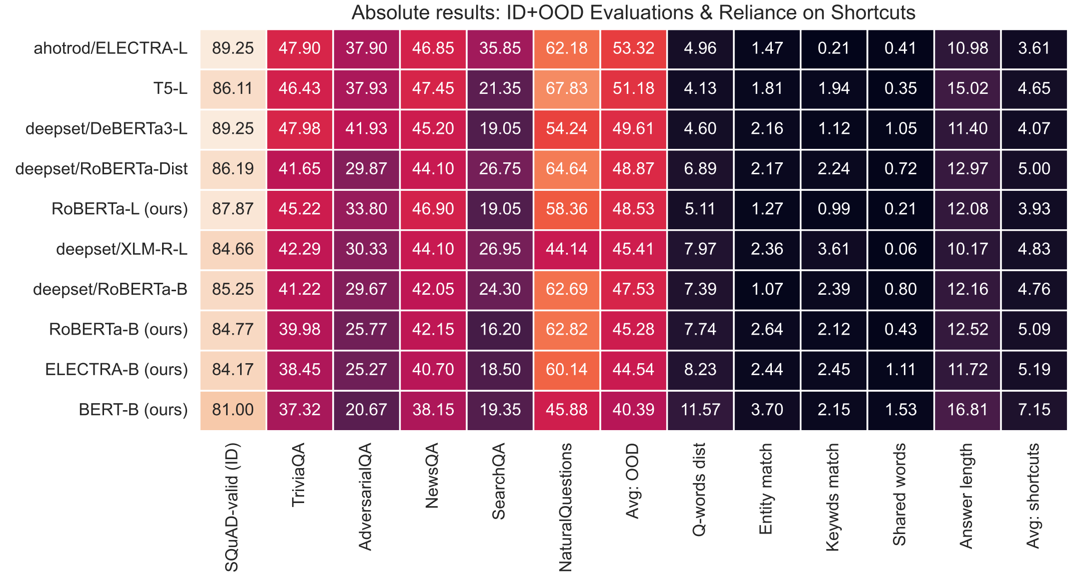

Figure 3: Exact-match results of (i) OOD evaluations and (ii) reliance on prediction shortcuts denoting a minimal relative drop in accuracy (in percentage) when facing data where the prediction shortcut is not applicable. See the *Prediction bias* algorithm in Mikula et al. (2024) for further details. These evaluations were used to compute the ranking and correlations in Figs. 1 and 2.

- **Keywords match** (Clark et al., 2019): Models' reliance on the matching keywords, i.e. low-frequency words between the question and answer.
- **Answer length** (Bartolo et al., 2020): Models' false reliance on a specific word-level length of the answer.
- **Entity match** (Mikula et al., 2024): Models' reliance on that the answer must contain a first entity matching the type of the question, such as "Who", "Where", etc.

B Experimental Details

Our experiments train a set of QA models, separately on each of our training and evaluation dataset. Towards the goal training models representative for real-world deployment, we perform hyperparameter search within each model family for optimal values of learning rate (including values 1e-6, 2e-6, 1e-5, 2e-5, 4e-5, 5e-5, 2e-4) and batch size (including values 8, 16, 32, 64). We used early stopping based on evaluation loss based on (in-distribution) validation set of SQuAD, patience=5, evaluations every 2000 updates and a maximum of 5 epochs. In this configuration, we were able to train each of our 11 trained models under 24 hours on a single Nvidia A40 GPU. Our training scripts are using HuggingFace Transformers library (Wolf et al., 2020). Training script can also be found among supplementary materials of this submission.

All our evaluations, including ID and OOD evaluations, employ exact-match metric and on the case of our own-trained models, we verify that the choice of our primary evaluation metric has no effect on the ranking of models in a majority of OOD evaluations, while causing at most two swaps in ranking in other cases.

Evaluations of Reliance on shortcuts follow the methodology of Mikula et al. (2024); here, we employ the `isbiased` library from the Authors' referenced GitHub repository[5]. Following the original methodology, we assess reliance on shortcuts consistently using the SQuAD 1.1's standard validation set of the full size. We include also our evaluation script among the supplementary materials of this submission.

B.1 Detailed Results

Figure 3 shows the detailed results with the absolute values of out-of-distribution evaluations as well as the relative dependencies on a reliance on surveyed spurious features for all our evaluated models. The OOD values are listed in exact-match metric, while the dependencies on spurious features are listed as a percentage of models' performance that *depends* on applicability of prediction shortcuts. Consistently with other figures, the models are ordered based on their absolute average dependency on shortcuts (last column).

[5] https://github.com/MIR-MU/isbiased

ahotrod/ELECTRA-L	ahotrod/electra_large_discriminator_squad2_512
deepset/DeBERTa3-L	deepset/deberta-v3-large-squad2
deepset/XLM-R-L	deepset/xlm-roberta-large-squad2
deepset/RoBERTa-B	deepset/roberta-base-squad2
deepset/RoBERTa-Dist	deepset/xlm-roberta-base-squad2-distilled

Table 2: Links to models on HuggingFace

SQuAD (Rajpurkar et al., 2016)	rajpurkar/squad
AdversarialQA (Jia and Liang, 2017)	UCLNLP/adversarial_qa
TriviaQA (Joshi et al., 2017)	mandarjoshi/trivia_qa
NewsQA (Trischler et al., 2017)	StellarMilk/newsqa
SearchQA (Dunn et al., 2017)	lucadiliello/searchqa
NaturalQuestions (Kwiatkowski et al., 2019)	sentence-transformers/natural-questions

Table 3: Links to datasets on HuggingFace

C Scientific artifacts used in this paper

We rely on publicly available datasets and models, all of them were made available under broad permissive licenses. At time of writing, none of the models we retrieved from HuggingFace have associated publications. To support a reproducibility of our results, we provide links to both existing models (Table 2) and datasets used in this work (Table 3).