On the Transferability of Minimal Prediction Preserving Inputs in Question Answering

Shayne Longpre* Apple Inc. slongpre@apple.com

Yi Lu* Apple Inc. ylu7@apple.com

Christopher DuBois Apple Inc. cdubois@apple.com

Abstract

Recent work (Feng et al., 2018) establishes the presence of short, uninterpretable input fragments that yield high confidence and accuracy in neural models. We refer to these as Minimal Prediction Preserving Inputs (MPPIs). In the context of question answering, we investigate competing hypotheses for the existence of MPPIs, including poor posterior calibration of neural models, lack of pretraining, and "dataset bias" (where a model learns to attend to spurious, non-generalizable cues in the training data). We discover a perplexing invariance of MPPIs to random training seed, model architecture, pretraining, and training domain. MPPIs demonstrate remarkable transferability across domains — achieving significantly higher performance than comparably short queries. Additionally, penalizing over-confidence on MPPIs fails to improve either generalization or adversarial robustness. These results suggest the interpretability of MPPIs is insufficient to characterize generalization capacity of these models. We hope this focused investigation encourages more systematic analysis of model behavior outside of the human interpretable distribution of examples.

1 Introduction

Feng et al. (2018) establish the presence of shortened input sequences that yield high confidence and accuracy for non-pretrained neural models. These Minimal Prediction Preserving Inputs (MPPIs) are constructed by iteratively removing the least important word from the query to obtain the shortest sequence for which the model's prediction remains unchanged (example shown in Figure 1).[1] Humans are unable to make either confident or accurate predictions on these inputs. Follow up work treats strong model performance on such partial-inputs as equivalent with models improperly learning the task (Feng et al., 2019; Kaushik and Lipton, 2018; He et al., 2019). Accordingly, we evaluate this proposition in question answering (QA), investigating the properties of MPPIs and how their existence relates to "dataset bias", out-of-domain generalization, and adversarial robustness.

SQuAD	
Context	... The site currently houses three cinemas, including the restored Classic the United Kingdom's last surviving news cinema still in full-time operation—alongside two new screens ...
Original	What's the name of the United Kingdom 's sole remaining news cinema ?
Reduced	news
Confidence	0.57 → 0.51

Figure 1: A SQuAD dev set example. Given the original *Context*, the model makes the same correct prediction ("Classic") on the *Reduced* question (MPPI) as the *Original*, with almost the same score. For humans, the reduced question, "news", is nonsensical.

First we examine the hypothesis that MPPIs are a symptom of poor neural calibration. Feng et al. (2018) propose we can "attribute [these neural] pathologies primarily to the lack of accurate uncertainty estimates in neural models." As neural models tend to overfit the log-likelihood objective by predicting low-entropy distributions (Guo et al., 2017) this can manifest in over-confidence on gibberish examples outside of the training distribution (Goodfellow et al., 2014). We test this hypothesis using pretrained models, shown to have better posterior calibration and out-of-distribution robustness (Hendrycks et al., 2020; Desai and Durrett, 2020). Contrary to expectations, we find large-scale pretraining does not produce more human interpretable MPPIs.

Second we examine the hypothesis that MPPIs are the symptom of "dataset bias" — where a

* equal contribution

[1] For question answering we construct MPPIs by only removing words from the query. Modifying the context paragraph is poorly defined in MPPI generation as it perturbs the output space, rendering an answer impossible or trivial.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1288–1300
June 6–11, 2021.

flawed annotation procedure results in hidden linguistic cues or "annotation artifacts" (Gururangan et al., 2018; Niven and Kao, 2019). Models trained on such data distribution can rely on simple heuristics rather than learning the task. As such, input fragments or "partial inputs" are often sufficient for a model to achieve strong performance on flawed datasets. This explanation has been considered for both Natural Language Inference tasks (the "hypothesis-only" input for Poliak et al. (2018); Gururangan et al. (2018)) and for Visual Question Answering (the "question-only" model for Goyal et al. (2017)). We expect models which rely on these spurious cues would fail to generalize well to other "domains" (datasets with different collection and annotation procedures). We discover even models trained in different domains perform nearly as well on MPPIs as on full inputs, contradicting this hypothesis. Further, we test their transferability across a number of other factors, including random training seed, model size, and pretraining strategy, and confirm their invariance to each of these.

Third we examine the hypothesis that MPPIs inhibit generalization. This intuition is based on MPPI's poor human interpretability, which could suggest models should not attend to these signals. To test this hypothesis, we regularize this phenomenon directly to promote more human understandable MPPIs, and measure the impact on out-domain generalization and adversarial robustness. Interestingly, out-domain generalization and robustness on Adversarial SQuAD (Jia and Liang, 2017) vary significantly by domain, with both declining slightly on average due to regularization.

In conjunction, these results suggest MPPIs may represent an unique phenomenon from what previous work has observed and analyzed. The performance of these inputs is not well explained by domain-specific biases, or posterior overconfidence on out-of-distribution inputs. Instead, this behavior may correspond to relevant signals as the impact of their partial mitigation suggests. We hope these results encourage researchers to not assume MPPIs, or other uninterpretable model behaviour, are dataset artifacts that require mitigation a priori. Before presenting mitigation solutions, we propose they follow a more systematic analysis proposed by our actionable framework by (a) rigorously testing the alleged causes of the observed behaviour, (b) confirming the bias does not generalize/transfer, and (c) ensuring the solution provides *consistent* improvements across domains within a task.

Dataset	ORIGINAL	BERT-B	XLNET-L
SQUAD (Rajpurkar et al., 2016)	11.54	2.32	2.65
HOTPOTQA (Yang et al., 2018)	18.96	2.07	2.55
NEWSQA (Trischler et al., 2016)	7.59	2.08	1.80
NATURALQ (Kwiatkowski et al., 2019)	9.17	1.22	1.26
TRIVIAQA (Joshi et al., 2017)	15.68	2.33	1.80
SEARCHQA (Dunn et al., 2017)	17.43	1.81	1.05

Table 1: Number of MPPI query tokens, for different datasets and models.

2 Experimental Methodology

All models trained, including DRQA (Chen et al., 2017), BERT (Devlin et al., 2019), and XLNET (Yang et al., 2019), employ setup and parameter choices from Longpre et al. (2019).[2] We generate MPPIs by iteratively removing the least important word from the question, while keeping the original prediction unchanged. The least important word is given as that for which the model's confidence in its prediction remains highest in its absence.[3]

To examine how MPPIs transfer across Question Answering domains we employ 6 diverse QA training sets and 12 evaluation sets.[4] The datasets were selected for annotation variety, differing on: question type, document source, annotation instructions, whether the question was collected independently of the passage, and skills required to answer the question. This set represents a realistic spectrum of domains for evaluating generalization.

We set aside $2k$ examples from each domain's validation sets in order to generate MPPIs for model evaluation. For each experiment we also generate a set of randomly shortened queries to compare against the MPPIs — we refer to this as the "Random MPPI" baseline. For each of the original examples, we generate this baseline by randomly removing words until the length matches that of the corresponding MPPI.

3 Experiments

3.1 Invariance of MPPIs

Feng et al. (2018) establish the "human-insufficiency" property of MPPIs for non-pretrained, LSTM and attention-based models, in-

[2]For DRQA, we borrowed the hyper-parameters from hitvoice (https://github.com/hitvoice/DrQA))

[3]Details of model training and examples of MPPI generation are described in Appendix A.

[4]Refer to Appendix A.3 for details, or the MRQA 2019 workshop: https://mrqa.github.io/shared. Fisch et al. (2019) normalized these datasets into purely answerable, extractive format.

	DrQA	BERT-B	XLNET-L
BERT-B	32.1 / 9.9	- / -	29.8 / 9.9
XLNET-L	26.0 / 7.2	29.8 / 9.9	- / -
RANDOM	13.0 / 1.8	12.6 / 0.9	14.2 / 1.3

Table 2: The mean similarity, measured in Jaccard Similarity / Exact Match (%), between the MPPIs from different model types and the random baseline.

cluding DrQA, and BiMPM (Wang et al., 2017). We extend this investigation for modern, pretrained Transformers, and assess the "invariance" of MPPIs: measuring whether they are random, or are affected by model architecture, pretraining strategy, or training dataset (domain).

In subsequent experiments we compare sets of MPPIs using the mean Exact Match or Generalized Jaccard Similarity (GJS), a variant of Jaccard Similarity, which accounts for the possibility of repeated tokens in either of the sequences being compared. Generalized Jaccard Similarity is defined between two token sequences X and Y in Equation 1. Here, n is the index of every element that appears in $X \cup Y$.

$$GJS(X,Y) = \frac{\sum_{i=1}^{n} min(X_i, Y_i)}{\sum_{i=1}^{n} max(X_i, Y_i)} \quad (1)$$

We will refer to this as "Jaccard Similarity" for simplicity.

3.1.1 Random Seed

First, we investigate whether MPPIs are "random", or influenced by weight initialization and training data order. Measuring the Jaccard Similarity between MPPI sequences produced by models with different training seeds we find $\text{JS}_{\text{MPPI}} = 57.1\% \pm 1.2$, as compared to $\text{JS}_R = 13.8\% \pm 0.8$ on the Random MPPI baseline. This suggests MPPIs are not simply the side-effect of randomness in the training procedure.

3.1.2 Pretraining and Architecture

One hypothesis is that traditional LSTM-based models, such as DrQA, do not have sufficient pretraining or "world knowledge" to rely on the entire sequence, and overfit to subsets of the input. If this were the primary source of MPPIs, we might expect models that are better calibrated and more robust to out-of-distribution examples to have longer and more interpretable MPPIs. Accordingly, we test this hypothesis with large pretrained transformers, which recent work demonstrates have better posterior calibration and robustness to out-of-distribution inputs.

Train Dataset	Reduction Dataset			
	SQuAD	HotPotQA	NewsQA	NatrualQ
SQuAD	- (-)	31.4 (8.8)	41.0 (21.6)	29.2 (12.5)
HotPotQA	39.7 (12.8)	- (-)	39.6 (18.8)	33.8 (13.5)
NewsQA	41.1 (13.0)	31.6 (7.2)	- (-)	35.2 (12.5)
NatrualQ	37.5 (12.7)	28.7 (7.1)	40.2 (17.9)	- (-)
Average	39.4 (12.8)	30.6 (7.7)	40.3 (19.4)	32.7 (12.8)

Table 3: The Jaccard Similarity (%) between BERT generated MPPIs, across domains. In parentheses are the Jaccard Similarity scores between the Random MPPI baseline and Train Dataset MPPIs.

Specifically, Desai and Durrett (2020) examine 3 separate NLP tasks, using "challenging out-of-domain settings, where models face more examples they should be uncertain about", and find that "when used out-of-the-box, pretrained models are calibrated in-domain, and compared to baselines, their calibration error out-of-domain can be as much as 3.5× lower". Similarly Hendrycks et al. (2020) systematically show "Pretrained transformers are also more effective at detecting anomalous or [out-of-distribution] examples". These findings suggest pretrained transformers should produce more interpretable MPPIs than non-pretrained models.

However, in Table 1 we show MPPIs remain incomprehensibly short for all 6 domains and even for pretrained transformer models (DrQA produces MPPIs on SQuAD of mean length 2.04). In Table 2 we show MPPIs produced by different model architectures and pretraining strategies are similar, significantly exceeding the Jaccard Similarity of the Random MPPI baseline ($\text{JS}_\text{R} = 13.8\%$). This would not be problematic if pretrained models produced lower confidences for MPPIs than the original examples (demonstrating some form of calibration). However, we find the opposite is true. Taking SQuAD for instance we see in 85% of cases the BERT model is more confident on the MPPI than the original example.

Lastly, we verify with manual grading tasks that the MPPIs for BERT and XLNet are no more interpretable to humans than DrQA's MPPIs, as shown in Table 5. This suggests that short, uninterpretable MPPIs are ubiquitous in modern neural question answering models and unmitigated by large scale pretraining, or improved out-of-distribution robustness.

3.1.3 Cross-Domain Similarity

Next, we investigate the extent to which MPPIs are domain-specific. We do this by measuring their similarity when produced by models trained in different domains. If MPPIs are the product of bias in the training data, such as annotation artifacts, we would expect them to be relatively domain specific, as different datasets carry different biases. In Table 3 a model trained from each domain (Train Dataset) generates MPPIs for each other domain (Reduction Dataset). For each Reduction Dataset, we measure the mean Jaccard Similarity between MPPIs produced by the Train Dataset model and MPPIs produced by the Reduction Dataset (in-domain) model. In parentheses we show the mean Jaccard Similarity between the Random MPPIs and the Train Dataset MPPIs. In all cases, MPPIs demonstrate higher similarity than the random baseline, indicating that they are not domain specific.

3.2 Cross-Domain Transferability of MPPIs

Even when models generate different MPPIs, they may still transfer to the other domain. We would like to measure MPPI transferability, independent of their similarity between models. If QA models perform well on MPPIs generated from a range of domains then this would suggest they are not a product of bias in the training data. Instead, they may retain information important to question answering, rather than annotation artifacts. To better measure the extent of MPPI transferability, we (a) train one model on SQuAD (Train Dataset), and another on NewsQA (Reduction Dataset), (b) use the NewsQA-model to generate $2k$ MPPIs on the NewsQA evaluation set, and (c) measure the F1 performance of the SQuAD-model evaluated on both the original NewsQA evaluation set and the MPPI queries as generated in part (b).

Figure 2 shows performance on out-domain MPPIs are 46.6% closer to original performance than on Random MPPIs. This evidence suggests MPPIs are highly transferable across domains. Consequently, MPPIs may relate to generalization, despite their poor human interpretability.

3.3 Human-Sufficient MPPIs do not Improve Generalization

Even though MPPIs are highly transferable between domains, their presence may be associated with poor generalization. To evaluate this possibility, we examine whether the penalization of MPPIs improves generalization, or adversarial robustness. While penalizing over-confidence on MPPIs has been shown to maintain equivalent in-domain performance, and yield subsequently longer and more human interpretable MPPI queries (Feng et al., 2018), its impact on generalization or robustness has not yet been examined.

Train Dataset	F1 Score (%)				
	△ID	OD	△OD	AR	△AR
SQuAD	-0.8	52.9	-1.5 ± 2.3	72.1	+3.1
HotpotQA	+0.6	48.5	-0.6 ± 1.2	45.5	+1.0
NewsQA	-0.9	53.0	-0.9 ± 0.6	62.9	-1.8
NaturalQ	+0.9	51.6	-2.9 ± 3.5	54.9	-0.9
TriviaQA	-0.6	42.3	-4.1 ± 2.8	38.9	-1.1
SearchQA	-0.5	38.0	-5.9 ± 2.9	32.3	-4.0
Overall Avg	-0.2	47.7	-2.7 ± 1.1	51.1	-0.6

Table 4: The impact of MPPI regularization on in-domain (ID) performance, macro-average out-domain (OD) generalization over 12 evaluation datasets, and adversarial robustness (AR) on Adversarial SQuAD. △X = F1 of MPPI regularized model minus F1 of regular model on target X (any of ID, OD, or AR).

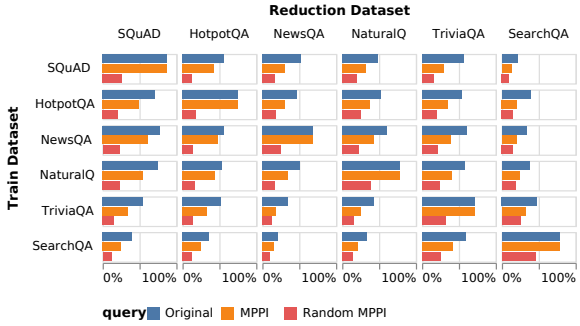

Figure 2: Cross-Domain Transferability: BERT question answering performance (F1) with different training sets (y-axis), and $2k$ evaluation sets (x-axis). Bars are colored by input type. On average, MPPI queries close the gap between models' performance on Random MPPIs and original queries by 46.6%.

We employ a simplified version of the MPPI penalization used by Feng et al. (2018), training a model with equal quantities of regular and MPPI examples — maintaining normal QA loss terms for the regular examples, and applying an entropy penalty to MPPI examples.[5] When penalizing over-confidence on MPPIs, we confirm the new MPPI length is significantly longer (Appendix sections B), and more human interpretable (Table 5).

In Table 4 we show the difference in F1 scores (△) between the regularized and original models. Results demonstrate that in-domain F1 (ID), macro-average out-domain F1 over 12 datasets (OD), and

[5] See Appendix section A.4 for details.

	HUMAN F1 (EM)
ORIGINAL QUERY	91.2 (82.3)
DRQA† MPPI	- (31.7)
BERT-B MPPI	41.6 (32.0)
XLNET-L MPPI	37.6 (26.0)
BERT-B* MPPI	60.7 (43.5)
RANDOM MPPI	26.5 (17.0)

Table 5: The mean human performance (in F1 and Exact Match over 100 examples) on different variants of MPPIs for SQUAD.
† Human performance cited from Feng et al. (2018)
* Indicates a model trained with MPPI regularization.

adversarial robustness F1 on Adversarial SQUAD (AR) all decline slightly on average with MPPI regularization — by 0.2%, 2.7%, and 0.6% respectively. These results suggest a model's ability to make predictions on MPPIs is not strongly correlated with either generalization or robustness across 13 total QA datasets. However, the relative stability of in-domain performance as compared to out-domain performance suggests mitigating MPPIs is more harmful when crossing domain boundaries.

Certain train datasets exhibit greater sensitivity to MPPI regularization than others. For instance SearchQA is drastically affected in all measures, HotpotQA hardly at all, and SQuAD actually improves by 3.1% in adversarial robustness. Additionally, Table 4 shows the 95% confidence intervals for out-domain generalization are often as large as the mean change in performance. Empirically, this demonstrates the effect of MPPI regularization is not consistent, having both positive and negative impacts on performance, depending on which of the 12 out-domain datasets is in question.[6]

4 Discussion

In SQUAD, the most common MPPI is the empty string (40%). Among non-empty strings, the most common MPPI tokens are: "what", "?", "who", "how", "when". Despite the pattern of interrogative words, these tokens are already among the most frequent in SQUAD questions, so it's challenging to measure the unique information they convey.

A more direct approach to understand the informative signal of MPPIs is to measure their "human insufficiency" property directly. We conduct a grading task, comparing human ability to answer real, MPPI, and random MPPI queries. Table 5 shows that humans could only correctly answer BERT and XLNet MPPIs slightly more often than random MPPIs (32% and 26% exact match compared to 17%), but could answer 43.5% of MPPIs produced by MPPI-regularized BERT. Although this confirms MPPI-regularization partially resolves over-confident behaviour for these human non-interpretable inputs, we've observed the resulting model fares slightly worse in domain generalization and robustness.

We find no evidence that MPPIs are explained by poorly calibrated neural models, lack of pretraining knowledge, or dataset-specific bias. Alternatively they may relate, at least in part, to useful and transferable signals. While practitioners, especially in model debiasing tasks, have focused on human understandable and generalizable features, this work would encourage them to also consider the presence of generalizable features which are not human interpretable. This observation closely relates to prior work in computer vision suggesting human uninterpretable, adversarial examples can be the result of "features", not "bugs", in which Ilyas et al. (2019) observe "a misalignment between the (human-specified) notion of robustness and the inherent geometry of the data." We hope this work provides a framework to rigorously evaluate the impact of bias mitigation methods on robustness and generalization, and encourages ML practitioners to examine assumptions regarding unexpected model behaviour on out-of-distribution inputs.

5 Conclusion

We empirically verify the surprising invariance of MPPIs to random seed, model architecture, and pretraining, as well as their wide transferability across domains. These results suggest that MPPIs may not be best explained by poorly calibrated neural estimates of confidence or dataset-specific bias. Examining their relationship to generalization and adversarial robustness, we highlight the ability to maintain in-domain performance but significantly alter out-domain performance and robustness. We hope our results encourage a more systematic analysis of hypotheses regarding model behavior outside the human interpretable distribution of examples.

6 Acknowledgments

We would like to acknowledge Eric Wallace, Shi Feng, Jordan Boyd-Graber, Christopher Clark, Drew Frank, Kanit Wongsuphasawat, Ni Lao, and Charlie Maalouf for their guiding insights and helpful discussion.

[6] See Figure 9 in Appendix A.4 for details.

References

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

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

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.

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 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*.

Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5533–5538, Florence, Italy. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334. IEEE.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

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

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

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.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. 2019. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pages 125–136.

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. Association for Computational Linguistics.

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*.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5010–5015, Brussels, Belgium. 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*, pages 4999–5007.

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.

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.

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

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

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.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

George Tsatsaronis, Michael Schroeder, Georgios Paliouras, Yannis Almirantis, Ion Androutsopoulos, Eric Gaussier, Patrick Gallinari, Thierry Artieres, Michael R Alvers, Matthias Zschunke, et al. 2012. Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Appendices

A Reproducibility

A.1 Question Answering Models

For reproducibility, we share our hyper-parameter selection in Table 6. We borrow our hyper-parameters from Longpre et al. (2019) for training all Question Answering (QA) models. Their parameters are tuned for the same datasets in the MRQA Shared Task. We found these choices to provide stable and strong results across all datasets.

Our BERT and XLNet question answering modules build upon the standard PyTorch (Paszke et al., 2019) implementations from HuggingFace, and are trained on 8 NVIDIA Tesla V100 GPUs.[7] For DrQA, by Chen et al. (2017), we borrowed the implementation and hyper-parameters from hitvoice (https://github.com/hitvoice/DrQA) and train on 1 NVIDIA Tesla V100 GPU.[8]

A.2 Dataset

We employ 6 diverse QA training sets and 12 evaluation sets from the MRQA 2019 workshop (https://github.com/mrqa/MRQA-Shared-Task-2019) (Fisch et al., 2019). These datasets have been normalized into purely extractive format and all questions are answerable. The 6 training datasets are SQuAD (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2016), TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), HotpotQA (Yang et al., 2018), and Natural Questions (Kwiatkowski et al., 2019). Six other evaluation datasets are included: BioASQ (Tsatsaronis et al., 2012), DROP (Dua et al., 2019), DuoRC (Saha et al., 2018), RACE (Lai et al., 2017), RelationsExtraction (Levy et al., 2017) , and TextbookQA (Kembhavi et al., 2017). Table 7 shows their statistics.

We use the hyperparameters described in Table 6 for training on each dataset. We use all the training data provided for each by MRQA.

A.3 Generating MPPIs

The process for generating MPPIs closely follows the procedure described by Feng et al. (2018). We operate with a beam size of $k = 3$, finding that larger beam sizes exhibit diminishing returns, and rarely produce different results. The procedure involves iteratively removing the token which is "least important" to the model. The least important token is defined as the one that when removed provides the smallest decrease in confidence in the originally predicted span. Note that in some cases confidence in the originally predicted span can even increase with the removal of a token. In any case, the least important token is always designated by the lowest confidence in the original prediction. The stop condition is when removing **any** additional token would change the model's prediction.

Note that we follow previous work in only removing words from the query in extractive question answering. The reason for this is the MPPI can be poorly defined when context tokens are removed. Since the output predictions are over the context tokens for extractive question answering, its possible to warp the answer space, or remove the answer

Model Parameter	Value
DrQA	
Model Size (# params)	$31.5M$
Avg. Train Time	$10h\ 30m$
Learning Rate	0.1
Optimizer	Adamax
Num Epochs	35
Batch size	32
Dropout	0.4
Hidden size	128
BERT-BASE	
Model Size (# params)	$108.3M$
Avg. Train Time	$2h\ 20m$
Learning Rate	$5e-5$
Optimizer	Adam
Num Epochs	2
Batch Size	25
Gradient Accumulation	1
Dropout	0.1
Lower Case	False
Max Query Length	64
Max Sequence Length	512
XLNet-LARGE	
Model Size (# params)	$364.5M$
Avg. Train Time	$4h\ 45m$
Learning Rate	$2e-5$
Optimizer	Adam
Num Epochs	2
Batch Size	6
Gradient Accumulation	3
Dropout	0.1
Lower Case	False
Max Query Length	64
Max Sequence Length	512

Table 6: Hyperparameter selection for each model type.

[7] https://github.com/huggingface/transformers

[8] We used the open source version available at https://github.com/hitvoice/drqa.

Dataset	Question (Q)	Context (C)	Avg. Q Len	Avg. C Len	Train	Dev
SQuAD	Crowdsourced	Wikipedia	11	137	86,588	10,507
NewsQA	Crowdsourced	News articles	8	599	74,160	4,212
TriviaQA	Trivia	Web snippets	16	784	61,688	7,785
SearchQA	Jeopardy	Web snippets	17	749	117,384	16,980
HotpotQA	Crowdsourced	Wikipedia	22	232	72,928	5,904
Natural Questions	Search logs	Wikipedia	9	153	104,071	12,836
BioASQ	Domain experts	Science articles	11	248	-	1,504
DROP	Crowdsourced	Wikipedia	11	243	-	1,503
DuoRC	Crowdsourced	Movie plots	9	681	-	1,501
RACE	Domain experts	Examinations	12	349	-	674
RelationExtraction	Synthetic	Wikipedia	9	30	-	2,948
TextbookQA	Domain experts	Textbook	11	657	-	1,503

Table 7: Statistics about datasets used: The first block presents six domains used for training, the second block presents six additional domains used for model evaluation and generating MPPIs.

altogether. Additionally, if we do not permit any alterations to the original prediction tokens, then there exists a trivial solution: remove all tokens except for the predicted answer. In this case an extractive question answering model is forced to predict that answer, with no alternative options. Consequently, MPPIs that allow modifications to the context, or output space, can be poorly defined. Since in question answering the query is an essential input to provide confident answers, we believe this is the most reasonable setup for the task.

$$p_{ij} = \max(\text{softmax}(S_i + E_j)) \tag{2}$$

For completeness, we describe our method of computing span confidence for question answering, given that there are many variations. Let $S \in \mathbb{R}^N$ be the vector of start logits and $E \in \mathbb{R}^N$ be the vector of end logits, both of sequence length N. For every combination of $i,\ j \in [0, N]$ where $j \geq i \leq min(j + C,\ N)$, and $C = 30$ is the maximum answer span length, we compute the confidence for that span of answer text as the sum of their respective logits $S_i + E_j$. The final confidence probability $p_i j$ for a given span is as shown in Equation 2.

The model, on the other hand, can still make the same prediction as it did on the full input, and with a similar degree of confidence.

A.4 Regularizing MPPIs

There are a couple differences between the MPPI entropy-regularization strategy employed in this work and in Feng et al. (2018). While Feng et al. (2018) fine-tune an a model already trained for the question answering task, we regularize MPPIs in the initial fine-tuning stage (starting from BERT and XLNet's pre-trained weights). Secondly, they alternate updates between two optimizers, one batch of maximum likelihood, two for MPPI entropy maximization, whereas we use the same optimizer and shuffle together equal numbers of MPPI and regular inputs. We find our method (without rigorous comparison) to be slightly more effective on BERT at mitigating the MPPI phenomenon (measured by subsequent MPPI length). We suspect, if there is an advantage, it is due to the regularization beginning with the start of fine-tuning, rather than on a subsequent stage of fine-tuning.

$$\mathcal{L}_{\mathcal{MPPI}} = C - \lambda \sum_{\tilde{x} \in \tilde{\mathcal{X}}} \mathbb{H}\left(f(y|\tilde{x})\right) \tag{3}$$

$$\mathcal{L} = \mathcal{L}_{\mathcal{QA}} + \mathcal{L}_{\mathcal{MPPI}} \tag{4}$$

For completeness, we provide our entropy regularization loss term in Equation 3. Let $\tilde{\mathcal{X}}$ denote the set of inputs that have been reduced to their MPPI, $\mathbb{H}(\cdot)$ denote the entropy and $f(y|x)$ denote the predicted confidence for y given x. We then represent the loss term for MPPIs as $\mathcal{L}_{\mathcal{MPPI}}$, where the constant $C = 10$ is chosen such that maximizing the entropy will minimize the loss. We use $\lambda = 0.1$ as the most effective choice in our limited set of trials. The full loss term, for all inter-mixed regular question answering, and MPPI examples is the sum of standard QA loss $\mathcal{L}_{\mathcal{QA}}$, and the MPPI loss term $\mathcal{L}_{\mathcal{MPPI}}$, as shown in Equation 4.

In Figure 3 we display the full comparison between the performances of the MPPI regularized models and the regular models on 13 QA datasets, including Adversarial SQuAD (Jia and Liang, 2017).

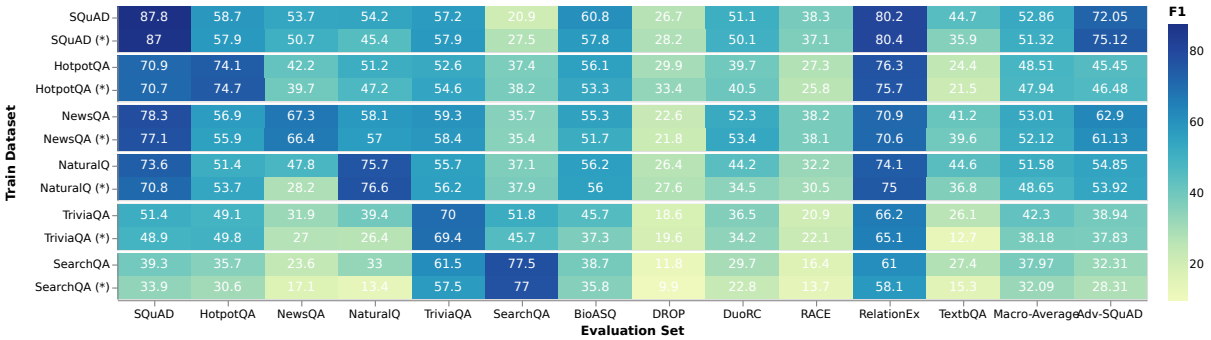

Figure 3: The generalization and robustness of BERT models evaluated on 12 datasets, as well as Adversarial SQuAD. The "(*)" indicates MPPI-regularization during training.

B How do MPPI Lengths Compare?

In the main paper we describe the differences in length distributions between original and MPPI queries. To provide more detail into the length distributions we plot histograms of the query word lengths, for the original queries, MPPI queries, and MPPI queries after the MPPI regularization procedure. These lengths are plotted below for SQuAD (Figure 4), HotpotQA (Figure 5), NewsQA (Figure 6), Natural Questions (Figure 7), TriviaQA (Figure 8), and SearchQA (Figure 9).

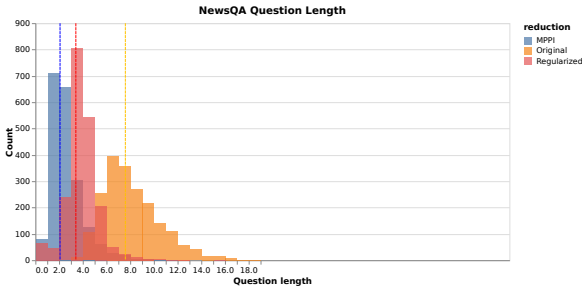

Figure 6: NEWSQA question length generated by different MPPI reduction methods

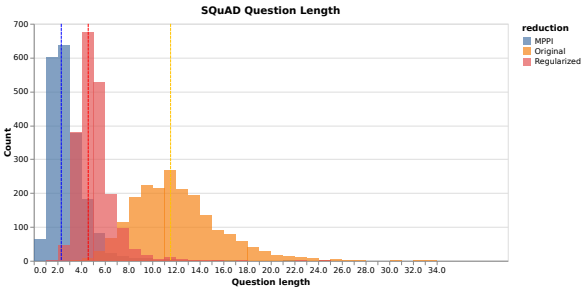

Figure 4: SQUAD question length generated by different MPPI reduction methods

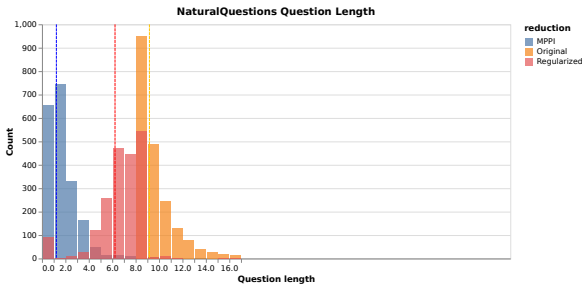

Figure 7: NATRUALQUESTION question length generated by different MPPI reduction methods

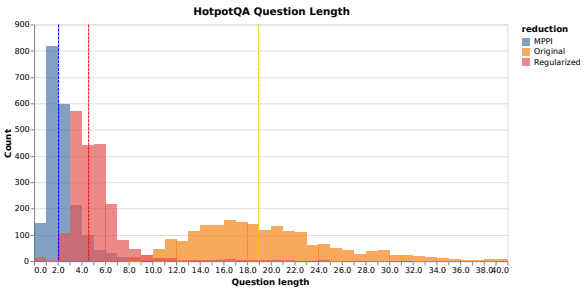

Figure 5: HOTPOTQA question length generated by different MPPI reduction methods

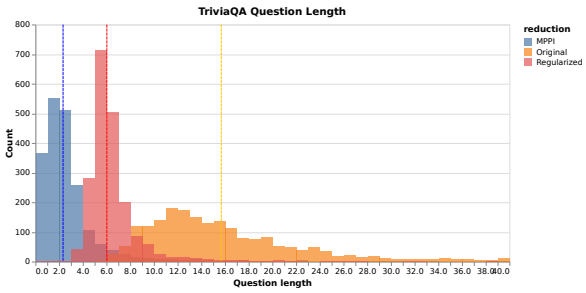

Figure 8: TRIVIAQA question length generated by different MPPI reduction methods

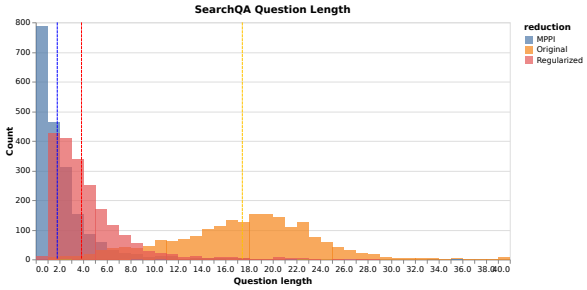

Figure 9: SEARCHQA question length generated by different MPPI reduction methods

The query length distributions show that MPPIs are significantly shorter than original queries, with the MPPIs of regularized models somewhere in between. These length distributions may be sufficient to explain why humans find the non-regularized MPPIs completely uninterpretable, and the regularized MPPIs somewhat more interpretable.

C Are MPPIs Invariant to Random Seed?

One of the preliminary questions in our investigation was whether changing the random training seed significantly altered the MPPI produced by a model. If it were the case that this had a drastic effect, we might suspect MPPIs were somewhat random, or the product of meangingless overconfidence on out-of-distribution inputs. Table 8 illustrates the random seed experiment in full. Training 10 SQuAD models, each with different random seeds, we generate MPPIs on the $2k$ SQuAD evaluation set, and compare 5 pairs. We measure the mean Generalized Jaccard Similarity of MPPIs produced by 2 models trained with different seeds.

We see the similarity between MPPIs trained with different seeds far exceed those of Rand-A, and Rand-B, which are akin to a "random" simulation of MPPIs. As with our previous random baselines these are generated by randomly sampling tokens from the original query, preserving word order, and ensuring that the length distribution matches that of the actual MPPIs to which they are being compared.

D Are MPPIs Invariant to Training Domain?

We discussed the invariance of MPPIs to training domain at length in the paper for BERT. For completeness, we provide the raw results for BERT in Table 9 and for XLNet in Table 10. These results show that MPPIs are far more similar to one another, even when training domain is different. The random baseline, in parenthesis, once again shows the Jaccard Similarity we would expect if MPPIs were purely random.

Seed A	Seed B	JS / EM
0	1	55.0 / 31.7
2	3	56.8 / 33.2
4	5	58.3 / 34.7
6	7	57.4 / 33.2
8	9	58.1 / 35.2
	Overall	57.1 / 33.6
Rand-A	Rand-B	13.8 / 0.9

Table 8: Observing the Jaccard Similarity and Exact Match between MPPIs on the SQuAD $2k$ evaluation set, we see significant token overlap despite seed differences. In contrast, the randomly generated sequences, preserving the length distribution of MPPIs, produces far less similar token sequences.

E Do QA Models Generalize to different MPPI Domains?

Expanding on the MPPI generalization analysis in Section 3.2, we provide the raw results. The cross-domain generalization of BERT and XLNet models on MPPIs sourced from different training domains is available in Table 11 and Table 12 respectively. Figure 10 visualizes how well XLNET generalizes to different MPPI domains. The results mirror those of BERT shown in the main paper.

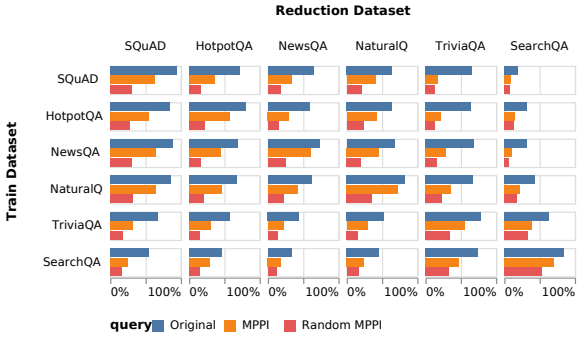

Figure 10: XLNET performance with different training sets (y-axis), and evaluation sets (x-axis). Bars measure the F1 score on the $2k$ evaluation set, colored by input type.

Train Dataset	Reduction Dataset					
	SQUAD	HOTPOTQA	NEWSQA	NATRUALQ	TRIVIAQA	SEARCHQA
SQUAD	- (-)	31.4 (8.8)	41.0 (21.6)	29.2 (12.5)	24.9 (10.9)	11.9 (9.6)
HOTPOTQA	39.7 (12.8)	- (-)	39.6 (18.8)	33.8 (13.5)	25.8 (10.7)	16.6 (12.6)
NEWSQA	41.1 (13.0)	31.6 (7.2)	- (-)	35.2 (12.5)	25.8 (10.8)	13.3 (9.4)
NATRUALQ	37.5 (12.7)	28.7 (7.1)	40.2 (17.9)	- (-)	25 (10.7)	15.0 (11.0)
TRIVIAQA	33.3 (13.0)	27.7 (8.0)	34.8 (18.1)	29.2 (15.4)	- (-)	23.1 (15.2)
SEARCHQA	23.4 (12.3)	16.8 (7.8)	24.7 (17.2)	24.4 (14.6)	23.4 (11.9)	- (-)
Average	35.0 (12.8)	27.2 (7.5)	36.1 (18.7)	30.4 (13.7)	25.0 (11.0)	16.0 (11.6)

Table 9: The Jaccard Similarity (%) between BERT generated MPPIs, across domains. The Random baseline MPPIs are in parentheses.

Train Dataset	Reduction Dataset					
	SQUAD	HOTPOTQA	NEWSQA	NATRUALQ	TRIVIAQA	SEARCHQA
SQUAD	- (-)	25.8 (9.0)	37.7 (19.7)	30.9 (11.1)	18.1 (10.5)	22.7 (26.3)
HOTPOTQA	28.4 (15.3)	- (-)	31.2 (17.6)	31.5 (12.4)	17.8 (12.0)	27.1 (25.5)
NEWSQA	31.8 (13.1)	25.3 (8.2)	- (-)	36.6 (11.9)	20.6 (9.0)	13.3 (11.7)
NATRUALQ	29.9 (12.9)	24.2 (8.4)	40.2 (16.8)	- (-)	22.3 (11.0)	19.2 (16.3)
TRIVIAQA	25.6 (14.8)	19.0 (8.0)	29.8 (17.4)	29.2 (13.7)	- (-)	31.2 (20.6)
SEARCHQA	21.6 (13.8)	15.5 (7.7)	25.2 (15.1)	24.6 (14.1)	28.3 (13.4)	- (-)
Average	27.5 (14.0)	22.0 (8.3)	32.8 (17.3)	30.6 (12.6)	21.4 (11.2)	22.7 (20.1)

Table 10: The Jaccard Similarity (%) between XLNET generated MPPIs, across domains. The Random baseline MPPIs are in parentheses.

Train Dataset	Query Type	SQuAD	HotpotQA	NewsQA	NaturalQ	TriviaQA	SearchQA
SQuAD	Original	87.74	56.31	48.81	21.53	56.74	52.62
SQuAD	MPPI	87.74	28.84	31.68	13.52	43.02	30.93
SQuAD	Random MPPI	26.42	16.19	19.69	9.55	13.01	17.46
TriviaQA	Original	54.64	71.04	42.27	47.53	51.85	34.45
TriviaQA	MPPI	34.25	71.04	24.67	32.28	33.85	18.91
TriviaQA	Random MPPI	15.21	32.23	15.92	25.49	14.12	12.67
NaturalQ	Original	75.28	58.18	77.78	37.84	54.08	51.16
NaturalQ	MPPI	55.15	40.43	77.78	24.32	44.52	34.25
NaturalQ	Random MPPI	23.28	23.94	38.43	18.39	16.32	17.34
SearchQA	Original	40.25	59.44	32.58	78.11	35.93	20.61
SearchQA	MPPI	24.84	41.67	21.24	78.11	24.96	15.9
SearchQA	Random MPPI	11.92	24.69	14.73	45.7	12.34	9.7
HotpotQA	Original	71.52	53.4	52.51	38.9	75.09	47.03
HotpotQA	MPPI	49.52	34.56	37.23	20.66	75.09	30.38
HotpotQA	Random MPPI	21.09	19.28	24.92	15.2	17.76	17.88
NewsQA	Original	78.16	60.91	59.94	33.79	56.53	68.19
NewsQA	MPPI	61.32	39.17	42.48	19.44	48.58	68.19
NewsQA	Random MPPI	22.78	20.4	22.74	15.05	14.49	24.83

Table 11: Cross-Domain Generalization of BERT Base models on different types of inputs. Values correspond to F1 scores on the question answering $2k$ evaluation set specified by the column.

Train Dataset	Query Type	SQuAD	HotpotQA	NewsQA	NaturalQ	TriviaQA	SearchQA
SQuAD	Original	93.92	64.97	66.62	15.13	70.51	65.06
SQuAD	MPPI	93.95	17.75	39.95	6.37	41.36	35.27
SQuAD	Random MPPI	31.0	12.86	21.12	7.3	16.19	17.85
TriviaQA	Original	67.55	78.15	51.7	67.69	57.85	44.97
TriviaQA	MPPI	31.77	78.18	27.8	43.63	32.34	22.24
TriviaQA	Random MPPI	17.01	34.03	16.03	33.18	14.54	13.27
NaturalQ	Original	85.61	67.84	82.06	42.92	67.43	60.82
NaturalQ	MPPI	63.02	35.73	82.06	20.44	46.19	42.95
NaturalQ	Random MPPI	31.74	23.31	36.07	18.05	19.98	22.45
SearchQA	Original	55.25	74.37	45.43	84.08	45.33	32.82
SearchQA	MPPI	25.6	47.5	26.33	84.08	29.41	18.57
SearchQA	Random MPPI	15.92	33.29	16.81	53.58	15.57	12.13
HotpotQA	Original	82.85	61.03	61.93	23.98	80.28	54.19
HotpotQA	MPPI	51.11	14.57	40.95	9.08	80.3	23.66
HotpotQA	Random MPPI	27.83	14.0	23.89	14.03	21.33	15.25
NewsQA	Original	88.56	69.32	67.61	30.74	69.14	73.17
NewsQA	MPPI	65.64	29.66	48.56	11.0	45.15	73.12
NewsQA	Random MPPI	30.47	16.29	20.27	6.64	15.71	25.35

Table 12: Cross-Domain Generalization of XLNET Large models on different types of inputs. Values correspond to F1 scores on the question answering $2k$ evaluation set specified by the column.