uFACT: Unfaithful Alien-Corpora Training for Semantically Consistent Data-to-Text Generation

Tisha Anders, Alexandru Coca, Bill Byrne
Department of Engineering, University of Cambridge, United Kingdom
anderstisha@gmail.com ac2123@cam.ac.uk wjb31@cam.ac.uk

Abstract

We propose UFACT (Un-Faithful Alien Corpora Training), a training corpus construction method for data-to-text (d2t) generation models. We show that d2t models trained on UFACT datasets generate utterances which represent the semantic content of the data sources more accurately compared to models trained on the target corpus alone. Our approach is to augment the training set of a given target corpus with *alien* corpora which have different semantic representations. We show that while it is important to have faithful data from the target corpus, the faithfulness of additional corpora only plays a minor role. Consequently, UFACT datasets can be constructed with large quantities of unfaithful data, minimising the need for faithful data. We show how UFACT can be leveraged to obtain state-of-the-art results on the WebNLG benchmark using METEOR as our performance metric. Furthermore, we investigate the sensitivity of the generation faithfulness to the training corpus structure using the PARENT metric, and provide a baseline for this metric on the WebNLG (Gardent et al., 2017) benchmark to facilitate comparisons with future work.

1 Introduction

Data-to-text (d2t) generation is the task of generating *fluent text* t given a set of information units, linearised into *data source* string d (Table 1).

d	{(name, Einstein), (born, 1879), (profession, physicist)}
t	Einstein was a physicist, born in 1879.

Table 1: Example of d2t system input (d) and output (t)

Training high quality generation models requires corpora whose reference texts are faithful to the data sources representing their semantic content, i.e. the reference texts t_r should have perfect information overlap with d. Most corpora are, however, noisy, with imperfect fact overlap between data d and reference text t_r (Dhingra et al., 2019a). The quality of the training data in that case negatively impacts the performance of a d2t generator trained on it, as well as making it difficult to estimate the true accuracy of a generation t_g, given t_r (Parikh et al., 2020). Faithful examples are however expensive to obtain, and usually only available in small quantities. In the context of this scarcity, we propose the UFACT training set construction method. UFACT allows a generator to learn a more accurate d2t generation model from a mixture of faithful and unfaithful corpora, which reduces the need for vast quantities of faithful examples. For instance, our best-performing UFACT dataset contains 88692 examples, of which only $20,000$ (24.34%) examples (the ones from the target corpus) are guaranteed to be faithful. We find that our approach leads to significant improvement in PARENT (Dhingra et al., 2019b) and METEOR (Banerjee and Lavie, 2005) compared to the conventional approach of training a d2t generator on one large unfaithful corpus. We conclude that even unfaithful examples from other corpora can contribute to fluency and faithfulness. Our UFACT-trained T5 surpasses state-of-the-art performance for METEOR on the WebNLG dataset.

2 Related work

Early approaches (Reiter and Dale, 1997) formalize d2t generation as three subtasks: content determination, structuring/grouping of information, and surface realisation. A handcrafted system is designed to solve each task. Recently, the focus has shifted towards end-to-end neural approaches, incorporating each of the subtasks into one system (Ferreira et al., 2019, Puduppully et al., 2018, Harkous et al., 2020).

A number of end-to-end approaches to increasing faithfulness in d2t generation are *curative*, i.e. address generation quality post-hoc. For instance, Harkous et al. (2020) and Dušek and Kasner (2020)

Findings of the Association for Computational Linguistics: ACL 2022, pages 2836 - 2841

produce candidate generations first, and then judge faithfulness with a separate model, by checking entailment between d and t_g. Another approach to enhance faithfulness is to alter the generation model. Chen et al. (2020b) propose a generation model comprised of a copy-generate gate within an LSTM positional encoder. The gate acts as a soft switch between a copy-from-data mode and a language-generation mode. Kale (2020) utilise transfer learning to enhance their generation model, through pre-training on a large unsupervised, task-agnostic corpus.

A different line of research focuses on *preventative* approaches, where the typical aim is to obtain a better model by improving the training data quality. Chen et al. (2020a) apply a unigram-based dataset selection process, by removing examples for which t_r is not sufficiently related to d. Parikh et al. (2020) also investigate this approach, releasing the noise-free ToTTo dataset, to ensure the training data does not encourage unfaithful generation. Filippova (2020) look for hallucinative examples in their dataset, either considering word-overlap, or comparing how strongly a language model vs. a conditional language model anticipates subsequent text. Dhingra et al. (2019b) develop the PARENT metric, a faithfulness-quantifying F-score that takes into account the data source in addition to the potentially divergent reference, providing a more robust assessment of the d2t mapping.

In their work on model-agnostic meta-learning, Finn et al. (2017) note that training on different instances of a required task (e.g., training on different corpora) can facilitate learning a particular task. Inspired by this approach, we add other corpora with different semantic representations to the training dataset. We find not only that adding corpora boosts the semantic faithfulness of the d2t generator, but also that said corpora need not necessarily satisfy stringent faithfulness requirements, unlike the target corpus.

3 Constructing a UFACT dataset

Typically, a d2t generation model is obtained by task-specific fine-tuning, where a large-scale pre-trained model such as T5 (Raffel et al., 2019) is fine-tuned on a small corpus. UFACT however, as an instance of *mixed-corpus training*, takes a different approach: examples from multiple corpora which do not share semantic representations, are linearised and tagged to form a large training corpus. A UFACT dataset is comprised of a *target* dataset for which we desire to maximise d2t generation fidelity and *alien* corpora. The latter are d2t corpora that may differ thematically and structurally from the target corpus and whose role is to improve generation fidelity on the target corpus.

3.1 Corpora included in the UFACT dataset

The UFACT datasets we experiment with are constructed from from three corpora which differ significantly in size, vocabulary, intended purpose, and linearisation technique. Figure 1 displays the relative sizes of the UFACT datasets (FU and FUU), their faithful counterparts (FF and FFF), as well as other dataset compositions examined.

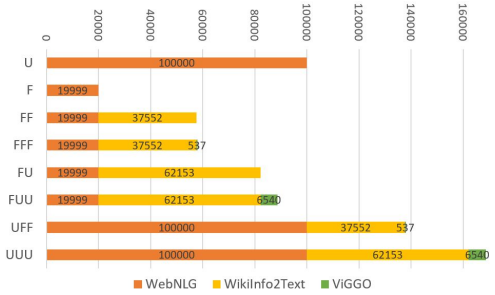

Figure 1: **Dataset sizes** The target corpus is WebNLG. Here U denotes *unfaithful*, describing a dataset that has not been curated while F stands for *faithful*, indicating a dataset that has been filtered to increase the faithfulness of the references to the data sources. See Appendix A for dataset curation approaches.

WebNLG examples consist of up to seven RDF-triplets (subject-predicate-object), which are atomic entities of a knowledge graph, linearised into a string. 15 topics appear, of which 10 are seen in training.

WikiInfo2Text[1] is based on slot-value pairs, imitating a table. Our WikiInfo2Text set (a subset of the original) comprises five topics (`UK_place`, `Book`, `Automobile`, `Military_conflict` & `French_commune`).

ViGGO (Juraska et al., 2019), a gaming dialogue corpus, has simple vocabulary, with 9 dialogue acts and 14 video game attributes available. The semantic representation consists of one dialogue act and 1-8 video game attributes, expressed as slot-value pairs that allow for lists of multiple values.

Table 2 shows a sample training point from each corpus. It also shows that in the joint dataset the data source of every example, d, is prepended with

[1] https://github.com/hitercs/WikiInfo2Text

d	**webnlg:** <s> Einstein <p> born <o> 1879 ; <s> Einstein <p> job <o> physicist
t	*Einstein was a physicist, born in 1879.*
d	**wikiinfo:** <name> H for Homicide && <author> S. Grafton && <series> Alpha Mysteries
t	*H for Homicide, by S. Grafton, is part of the Alpha Mysteries series.*
d	**viggo:** <request_explanation> (<rating>:[excellent], <genres>:[shooter, RTS]>)
t	*What is it about shooter and RTS games that you find so great?*

Table 2: **Examples of the three d2t corpora.** *WebNLG* consists of subject-predicate-object triplets, marked as such with `<s>`, `<p>`, `<o>`. *WikiInfo2Text* has slot-value pairs, with slot-names in angle brackets, and pairs separated by `&&`. *ViGGO* has limited vocabulary, but the hierarchical structure of a dialogue act (e.g., `request_explanation`) parametrized by slot-value pairs (e.g., `<rating>:[excellent]`).

a dataset-specific tag (`webnlg:`, `wikiinfo:`, `viggo:`). Tags are usually task-based, (e.g., `translate eng-to-ger:`) and have been shown to be particularly effective with Transformer models (Ribeiro et al., 2021). Treating each dataset as a different instance of the d2t task as in the meta-learning approach, the tags reveal an example's affiliation with a dataset.

3.2 Assembling a UFACT dataset

In summary, a UFACT dataset is a mixed corpus comprising a *target* (WebNLG) and *alien* datasets (WikiInfo2Text & ViGGO). The next section shows that while the target corpus should obey a maximum degree of faithfulness, the faithfulness of alien datasets plays a subordinate role. Therefore, in a UFACT dataset, the target corpus obeys the *quality-over-quantity* principle, whereas alien corpora prioritise *quantity over quality*.

4 Experiments

4.1 Experimental setup

We fine-tune the pre-trained T5-base (Raffel et al., 2020) from HuggingFace[2] for one epoch with batch size 8. We report averages of 5 values, obtained from training the model with 5 different seeds. We measure METEOR, BLEU (up to 4-grams) and PARENT (Dhingra et al., 2019b), a metric specifically developed for d2t-generation, considering both the reference text and the data source. PARENT uniquely assesses the faithfulness of the generation to the data source. For computing PARENT, we use both the word-overlap (P(w)) and co-occurrence (P(c)) entailment models. All models are tested on the WebNLG test set, as in Harkous et al. (2020), to provide a fair comparison. The dataset compositions for different experiments are given in Figure 1.

[2] https://huggingface.co/t5-base

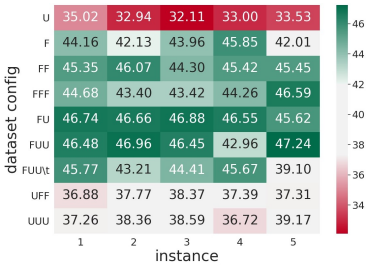

Figure 2: **T5 instance PARENT scores** for each model instance (i.e. data configuration). 'FUU\t' is a UFACT dataset without tags.

4.2 Effect of training dataset structure

Table 3 and Figure 2 show the effect of the training set structure on the model performance.

	Web.	Wik.	ViG.	P(w)↑	P(c)↑	M↑	B↑
1	U	-	-	33.32	44.43	48.28	18.89
2	F	-	-	**43.62**	**55.57**	**60.28**	**42.03**
3	F	F	-	**45.32**	**58.19**	**61.36**	39.1
4	F	F	F	44.47	56.17	60.13	**40.61**
5	F	U	-	**46.49**	**58.95**	**61.81**	**41.48**
6	F	U	U	46.02	58.54	61.59	40.88
7	F\t	U\t	U\t	43.63	59.32	60.06	33.71
8	U	F	F	37.54	48.70	51.02	**25.16**
9	U	U	U	**38.07**	**51.04**	**52.31**	18.85

Table 3: **Experimental results for T5, with different dataset configurations.** PARENT, METEOR and BLEU scores are measured for dataset configurations involving WebNLG (target), WikiInfo2Test (alien) & ViGGO (alien), respectively.{F,U}\t=no tags. All numbers reported are averages of the score of 5 models.

Training on single datasets *(Table 3, rows 1-2)* When training on the target dataset alone (i.e., WebNLG) a large performance boost is obtained on all metrics from using the faithful dataset WebNLG[F], despite the fact that it contains only 20% of the examples in WebNLG[U] (Figure 1). This demonstrates the detrimental effect of unfaithful target datasets, which are commonly used, on d2t generation faithfulness. The METEOR score of 48.28 on WebNLG[U] is comparable to the range of $\sim$ 39 - 46 reported in previous work (Ribeiro et al., 2021). Using faithful in-domain data has a large positive effect on all metrics (row 2).

Addition of faithful alien corpora *(rows 3-4)* When augmenting the target corpus with faithful

alien corpora (i.e. F-F & F-F-F), the training corpus size increases by factors of 1.88 and 1.90, respectively. As expected, performance increases on PARENT and METEOR, compared to faithful single-corpus training (F). However, F-F (i.e. just one alien dataset) outperforms F-F-F (two alien datasets). This may be due to the fact that ViGGO has a complex semantic representation diverging from the tuple/triplet representation in the other datasets, differs considerably in domain[3] from WebNLG and WikiInfo2Text, and only represents 0.92% of the F-F-F dataset (Figure 1). Therefore, it may act as too strong a regulariser during the training phase. The decrease in BLEU coupled with increases in METEOR and PARENT suggests that the generation model stays more faithful to the table, while also phrasing the sentence in its own way.

Training on UFACT datasets *(rows 5-6)* Training on UFACT datasets F-U and F-U-U improves generator perfomance compared to training with the faithful counterparts (F-F & F-F-F) (rows 3-4). This increase shows that the faithfulness of alien datasets WikiInfo2Text and ViGGO plays a subordinate role, and the model instead benefits from the sheer number of fluent examples. However, with the addition of ViGGO[U] (row 6 vs. row 5), no metric score is boosted, suggesting a constraint on alien datasets in terms of how much domains and, potentially, semantic representation can differ.

UFACT without tags *(row 7)* Training on the largest mixed corpus (F-U-U) without dataset-specific tags reduces every metric's score, with the exception of P(c) which increases by 1.33%. Coupled with the decrease in P(w) and BLEU this suggests that the generated text contains less lexical overlap with the references.

Can the target corpus be unfaithful? *(rows 8-9)* We have seen that the large unfaithful target corpus WebNLG[U] alone is the worst-performing dataset configuration. The addition of alien corpora in this case, unlike in previous experiments, does not lead to state-of-the-art-like performance. Metric scores stay significantly below any dataset with a faithful target corpus, including the UFACT datasets. The low performance in unfaithful-target-corpus configurations shows that the straightforward addition of alien corpora does not automatically result in desirable scores, and therefore justifies UFACT's quality-over-quantity principle for the target corpus.

[3]ViGGO has gaming-related chatbot-like utterances, whereas WebNLG and WikiInfo2Text center around geography, history, culture and public life.

4.3 Analysis of UFACT efficacy

The above results indicate that faithfulness in the target corpus should not be compromised, not even to gain a larger training set (see largest dataset U-U-U vs. smallest dataset F, or simply F vs. U). Furthermore, faithful alien corpora cannot compensate for unfaithful target corpora (e.g. U-F-F vs. F).

While faithful examples are also desirable in alien datasets, the trade-off between performance and effort for faithful examples is such that faithfulness is not worth pursuing at any cost, seeing that F-U / F-U-U outperform F-F / F-F-F.

The UFACT-method however insists on the target corpus being faithful.

Models trained with $N = 2$ corpora outperform those with $N = 3$ in this paper, suggesting that adding corpora with significantly different domain coverage and semantic representations may be counterproductive when those corpora make up a tiny portion of the dataset. Subsequently, the regularising effect is mitigated in F-U-U, since the portion of ViGGO is higher (7.37%).

Both METEOR, a reference-based metric and PARENT(c/w), which both take the reference *and* the data source into account, increase when training on UFACT datasets compared to conventional training (row 6 vs. 1). These increases suggest the data source is more accurately represented in the generated text. Therefore, UFACT provides a method of training better d2t models, with increased semantic faithfulness. The efficacy of mixed-corpus training shows that pretrained language models are powerful enough to learn and benefit from several tasks at once, provided the tasks are similar enough and sufficiently represented among the training set.

On WebNLG, UFACT achieves a new state-of-the-art result of 61.81 on METEOR (Ribeiro et al., 2021) (Table 4).

Author	Model/Method	M	B
Castro Ferreira et al. (2019)	UPF-FORGe	39.00	38.65
Harkous et al. (2020)	DATATUNER	42.40	52.90
Kale (2020)	T5-large	44.00	**61.44**
Moryossef et al. (2019)	StrongNeural	39.20	46.5
Schmitt et al. (2020)	Graformer	43.38	61.15
Zhao et al. (2020)	PLANENC	41.00	52.78
our paper	UFACT	**61.81**	41.84

Table 4: **State-of-the-art results on WebNLG** for METEOR and BLEU.

The comparatively low BLEU scores, in combination with high METEOR scores, are arguably desirable, since n-gram precision metric BLEU rewards simply copying from potentially unfaithful t_r, whereas METEOR can also reward semantically equivalent rephrasings of t_r. METEOR and BLEU results thus suggest high semantic overlap without copying. Meanwhile, UFACT datasets F-U-U and F-U achieve the highest PARENT scores (Table 3, rows 5-6), ensuring semantic overlap with both reference *and* data source.

5 Conclusion

We have presented the UFACT-method, which boosts the faithfulness of data-to-text generation models by appropriately constructing the training corpus. Training T5 on a mixture of d2t corpora results in strong semantic accuracy increase, as long as and the target corpus remains faithful. UFACT's lax constraints on the majority of the training set mitigates the scarcity problem in finding faithful d2t corpora, thus making faithful d2t generation more practically feasible. The new state-of-the-art METEOR score proves that language models alone, if trained with a carefully constructed dataset, can be highly effective data-to-text generators.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. 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 552–562, Hong Kong, China. Association for Computational Linguistics.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020a. KGPT: knowledge-grounded pre-training for data-to-text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8635–8648. Association for Computational Linguistics.

Zhiyu Chen, Harini Eavani, Wenhu Chen, Yinyin Liu, and William Yang Wang. 2020b. Few-shot NLG with pre-trained language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Online. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019a. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur P. Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. 2019b. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4884–4895. Association for Computational Linguistics.

Ondřej Dušek and Zdeněk Kasner. 2020. Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. *CoRR*, abs/1908.09022.

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *CoRR*, abs/2010.05873.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135. PMLR.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation, INLG 2017, Santiago de Compostela, Spain, September 4-7, 2017*, pages 124–133. Association for Computational Linguistics.

Hamza Harkous, Isabel Groves, and Amir Saffari. 2020. Have your text and use it too! end-to-end neural data-to-text generation with semantic fidelity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2410–2424, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Juraj Juraska, Kevin Bowden, and Marilyn Walker. 2019. ViGGO: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 164–172, Tokyo, Japan. Association for Computational Linguistics.

Mihir Kale. 2020. Text-to-text pre-training for data-to-text tasks. *CoRR*, abs/2005.10433.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. 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 2267–2277, Minneapolis, Minnesota. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2018. Data-to-text generation with content selection and planning. *CoRR*, abs/1809.00582.

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. *CoRR*, abs/1910.10683.

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

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87.

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating pretrained language models for graph-to-text generation. *ArXiv*, abs/2007.08426.

Martin Schmitt, Leonardo F. R. Ribeiro, Philipp Dufter, Iryna Gurevych, and Hinrich Schütze. 2020. Modeling graph structure via relative position for better text generation from knowledge graphs. *CoRR*, abs/2006.09242.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

A Obtaining faithful versions of the corpora

A.1 WebNLG & ViGGO

For WebNLG and ViGGO, faithful examples were retrieved from Harkous et al. (2020)[4], by selecting semantic fidelity classifier training examples labelled `accurate`.

A.2 WikiInfo2Text

Slot-value pairs with slot names which are by default irrelevant to the text (*e.g.* `img_size`, or other website-specific meta-data) were excluded from the respective example.
To be included in the training dataset, WikiInfo2Text examples had to obey two hand-crafted rules:

1. Generation-to-data-source length ratio:
   - To prevent references from giving information beyond the data source, the number of characters in the generation was restricted, given the number of semantic components in the data source:
     `len(ref) < tau*num_datapts`
2. Overall reference text length:
   - To avoid hallucinative reference texts, the number of characters in the reference was restricted:
     `len(ref) < lambda`

Values for τ and λ can be found in the table below. For WikiInfo2Text, we still perform some superficial cleaning to prevent extremely long examples from overloading the GPU.

	τ	λ
WikiInfo2Text[F]	60	800
WikiInfo2Text[U]	150	1500

Table 5: WikiInfo2Text cleaning parameter settings

[4] https://github.com/amazon-research/datatuner/tree/main/paper