CoDa: Constrained Generation based Data Augmentation for Low-Resource NLP

Chandra Kiran Evuru[♠*] **Sreyan Ghosh**[♠*] **Sonal Kumar**[♠] **Ramaneswaran S**[♣]
Utkarsh Tyagi[♠] **Dinesh Manocha**[♠]
[♠]University of Maryland, College Park, USA [♣]NVIDIA, Bangalore, India
{ckevuru, sreyang, sonalkum, utkarsht, dmanocha}@umd.edu, ramanr@nvidia.com

Abstract

We present **CoDa** (**Co**nstrained Generation based **Da**ta Augmentation), a controllable, effective, and *training-free* data augmentation technique for low-resource (data-scarce) NLP. Our approach is based on prompting off-the-shelf instruction-following Large Language Models (LLMs) for generating text that satisfies a set of constraints. Precisely, we extract a set of simple constraints from every instance in the low-resource dataset and verbalize them to prompt an LLM to generate novel and diverse training instances. Our findings reveal that synthetic data that follows simple constraints in the downstream dataset act as highly effective augmentations, and CoDa can achieve this without intricate decoding-time constrained generation techniques or fine-tuning with complex algorithms that eventually make the model biased toward the small number of training instances. Additionally, CoDa is the first framework that provides users explicit control over the augmentation generation process, thereby also allowing easy adaptation to several domains. We demonstrate the effectiveness of CoDa across 11 datasets spanning 3 tasks and 3 low-resource settings. CoDa outperforms all our baselines, qualitatively and quantitatively, with improvements of 0.12%-7.19%. Code is available [1].

Method	**Original 1:** Scott Gimple sets the record straight on that big midseason finale scene. **Original 2:** mine is all of Isaiah chapter 6 because that is what Jesus did for me.
BackTrans (Yu et al., 2018)	1. Scott Gimple is setting the record on this. 2: all of Isaiah, chapter 6, is mine, for this is what Jesus.
EDA (Wei and Zou, 2019)	1: scott gimple sets the record straight on that big midseason finale scenery. 2: mine is all of isaiah chapter because personify that is what jesus did for me.
SSMBA (Ng et al., 2020)	1. scott gimple sets the record, on that big preseason night website. 2: it in all of isaiah - 6 because that is what you did for me.
GPT3Mix (Yoo et al., 2021)	1: Scott Gimple sets the record straight on that big midseason finale scene. 2: I cherish Isaiah chapter 6 as it embodies what Jesus has done for me.
GENIUS (Guo et al., 2022)	1: For the record, there was no midseason finale scene. That is, until the finale. 2: Isaiah chapter 6 because it says, "If you don't believe in God, you will die."
CoDa *(ours)*	1: The recording of the scene in the new movie by Scott Gimple was a big success, capturing the essence of the entertaining story. 2: Jesus taught us to be free, to follow our hearts and minds, and to live life to the fullest. In today's society, we must continue to mine the teachings of Jesus to find the courage to live our lives on our own terms.

Table 1: Comparison of augmentations generated using CoDa with our baselines. CoDa generates augmentations that are more coherent and diverse. More examples in Appendix E.2.

1 Introduction

Data augmentation is a widely used technique to address the problem of limited training data in low-resource NLP (Chen et al., 2023). Owing to the recent advancement in generative AI, using synthetic data to train task-specific models has also gained much popularity. However, given a low-resource NLU dataset, effectively generating task-specific data to expand the dataset still poses a significant challenge. For example, while diversity in tokens and contexts in generated augmentations typically benefits downstream performance, excessively diverse examples may negatively impact consistency with the underlying downstream data distribution, thereby hurting performance (Geiping et al., 2023). This highlights the importance of having more control during the generation process to ensure data augmentation is done effectively.

In the past, researchers have employed methods like text-editing (Wei and Zou, 2019; Karimi et al., 2021; Shou et al., 2022), fine-tuning Pretrained Language Models (PLMs) with various algorithms (Wang et al., 2022; Zhou et al., 2021; Guo et al., 2022; Ghosh et al., 2023a,c), etc. However, most of these methods do not impose explicit controls to achieve diversity or consistency. The recent rise of autoregressive LLMs, known for their advanced generative and reasoning skills, introduces promising yet under-explored opportunities to enhance diversity in task-specific synthetic data synthesis. However, controlling autoregressive generation has proved to be innately challenging and complex (Zhang et al., 2023), and prompting-based methods have often employed manual human efforts for extracting data attributes that promote consistency (Yu et al., 2023).

[1]https://github.com/Sreyan88/CoDa
[*]These authors contributed equally to this work.

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3754–3769
June 16-21, 2024 ©2024 Association for Computational Linguistics

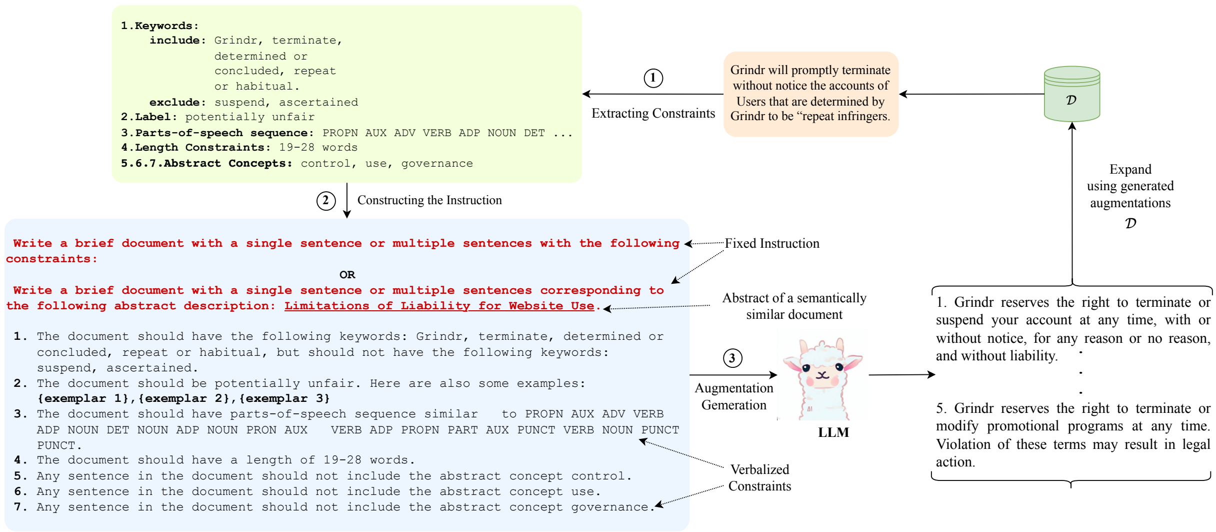

Figure 1: Illustration of **CoDa**. ① For every document in a low-resource NLU dataset $\mathcal{D}$, we extract a set of simple heuristic-based constraints from and ② verbalize them to generate an instruction. ③ This instruction is then fed to an existing instruction-tuned LLM for generating augmentations, which are then added to $\mathcal{D}$ for training a downstream model.

Main Contributions. We propose CoDa, a novel and effective data augmentation methodology for low-resource NLP. CoDa works with any off-the-shelf instruction-tuned LLM in a training-free fashion and provides explicit control over generated augmentations. We first extract simple heuristic-based constraints from training instances in a low-resource NLU dataset and then verbalize them to construct a natural language instruction. Next, we use this instruction to prompt an LLM for generating augmentations (example in Fig. 1). Alternative to complex decoding-time-constrained generation methods and manual attribute extraction, CoDa provides a simpler and more intuitive natural language-based interface for constrained generation. CoDa is also the first framework to explore controlled generation for data augmentations, which ensures that the synthetic data is closely aligned with the specific needs of the task and characteristics of the target domain. We show that CoDa, which is *training-free* and much simpler, quantitatively and qualitatively outperforms all prior-art by 0.12%-7.19% across various settings.

2 Related Work

Generative data augmentation for low-resource NLP has been extensively studied in prior work and can be categorized into four primary techniques. Firstly, **text-infilling** involves corrupting source text segments and using a PLM to refill these gaps. This process often relies on conditioning the corrupted text, a concept also known as keyword conditioning in some studies (Zhou et al., 2021; Guo et al., 2022; Ghosh et al., 2023c,a,b). Secondly, **text editing** focuses on modifying certain parts of a given sentence (Wei and Zou, 2019; Shou et al., 2022). Thirdly, **prompting** involves generating new training sentences by prompting LLMs (Ye et al., 2022; Sahu et al., 2023), which can be further uncategorized into conditioning attributes, exemplars, or constraints derived from training data.

3 Methodology

Fig. 1 illustrates the CoDa pipeline. Given a low-resource dataset $\mathcal{D} = \{d_0, \cdots, d_i, \cdots d_n\}$, we first extract a set of simple heuristic-based constraints from each document d_i and then verbalize the constraints to construct an instruction I_{d_i}. After this, we either instruct the LLM with I_{d_i} to generate a completely new document or rephrase another existing document from $\mathcal{D}$. For the latter, we first retrieve a document from $\mathcal{D}$, convert it into its short and concise abstract description by prompting an LLM, and then employ I_{d_i} to generate a document from the abstract description and the extracted constraints. For retrieval, we calculate cosine similarity between SentenceBERT embeddings (Reimers and Gurevych, 2019) of the source document d_i and all other documents in $\mathcal{D}$, and we randomly sample a sentence from the top-k and bottom-k similar sentences. For a total of 5 augmentations, we generated 3 novel documents and rephrased 2 other documents for every d_i. Finally, all the generated augmentations are added to $\mathcal{D}$ for training a downstream NLU model. We now describe our methodology to extract constraints in detail.

3.1 Extracting Constraints

a) Lexical Constraints. Inspired by a wealth of prior work in generative data augmentation and constrained generation (Zhou et al., 2023), we extract a set of keywords from a source sentence and constrain the augmentations to contain these keywords. More specifically, given a source document d, we first extract all its *n*-grams (1 to 3-grams) $\mathrm{N} = \{\mathrm{n}_0, \cdots, \mathrm{n}_\mathrm{t}, \cdots, \mathrm{n}_\mathrm{T}\}$. Next, we assign an importance score to each by calculating cosine similarity between $\mathbf{E}(\mathrm{n}_\mathrm{t})$ and $\mathbf{E}(\mathrm{d})$, where $\mathbf{E}$ is pretrained SentenceBERT. Finally, we select the top-*k* *n*-grams as our keywords. Additionally, for tasks like NER and QA, we add the corresponding target spans to the list.

b) Syntactic Constraints. In formal domains such as legal and biomedical, language is often governed by syntactical structures. Following a predefined POS pattern ensures that the generated sentences adhere to the formal style and tone expected in the domain. Readers can refer to Appendix 10 for some examples. Thus, we consider syntactic constraints that necessitate the generated augmentations to adhere to specific syntactic rules. More specifically, we extract the part-of-speech sequence from a randomly chosen sentence in d and constrain our generations to adhere to the sequence for a particular sentence.

c) Semantic (Label) Constraints. A primary requirement for effective data augmentations is that the semantics of the generated augmentations adhere to the underlying label of the source document d. To satisfy this, we consider label constraints so that the generated augmentations align closely to the original target label (e.g., *positive* sentiment). We use the target label of d with 3 exemplars for this constraint. The exemplars are chosen randomly from the dataset $\mathcal{D}$ and placed in random order in the final instruction.

d) Length Constraints. Length mismatches between training and testing instances have been known to degrade downstream NLU performance (Rogers et al., 2021). Motivated by this, we consider length constraints that necessitate the total number of tokens in the generated augmentations to fall within a specified range. We calculate the total number of tokens in d and add and subtract sd from it to obtain the lower and upper limits of the range, respectively. The value of sd is determined by computing the standard deviation of length distribution across the entire dataset $\mathcal{D}$.

e) Concept Constraints. The presence of spurious features in the training set causes the downstream NLU model to adopt shortcut learning strategies, impacting its performance in real-world, atypical situations where these features are not present (Sagawa* et al., 2020). Data augmentations can further amplify such spurious features in $\mathcal{D}$ if not handled correctly. We propose a novel strategy to ensure that generated augmentations do not have spurious features. We first employ the method proposed by Friedman et al. (2022) to extract a list of spurious phrases for each label in the dataset. We then pass these phrases with example sentences consisting of these phrases to an LLM and ask it to return a short abstract concept that the spurious phrases describe in the documents (e.g., *rating in movie reviews* for negative reviews in the IMDB dataset). Finally, we select the top 3 abstract concepts for each label and add is as a negation constraint for augmentation generation.

3.2 Constructing the Instruction

After extracting the constraints from d, we verbalize the constraints to a single instruction for prompting an instruction-tuned LLM. The verbalization is done through fixed hand-written templates. An example of an instruction is shown in Fig. 1.

4 Experimental Setup

Baselines. Gold-only refers to training our model only on the low-resource gold data. For sequence classification (SC), we compare CoDa with text editing baselines: EDA (Wei and Zou, 2019), AEDA (Karimi et al., 2021), and AMR-DA (Shou et al., 2022), learning-based infilling baselines: SSMBA (Ng et al., 2020), GENIUS(-**ft**) (Guo et al., 2022), PromDA (Wang et al., 2022), LLM-based prompting baselines: ZeroGen (Ye et al., 2022), GPT3Mix (Yoo et al., 2021) and rephrasing baselines: BackTrans (Yu et al., 2018). For the Intent Classification task, specifically in SC, we add another LLM-based prompting baseline: PromptMix (Sahu et al., 2023). For Named Entity Recognition (NER), we compare CoDa with LwTR (Dai and Adel, 2020), DAGA (Ding et al., 2020), MELM (Zhou et al., 2021), PromDA (Wang et al., 2022) and ACLM (Ghosh et al., 2023c). Finally, for question answering (QA), we compare it with ZeroGen, BackTrans, GENIUS, EDA, and AEDA. Details on the working of all baselines are provided in Section D.

Model	Huffpost			Yahoo			OTS			ATIS			Massive		
	100	200	500	100	200	500	100	200	500	100	200	500	100	200	500
Gold	76.82	77.96	80.51	42.50	49.50	55.47	74.75	83.49	**95.14**	85.13	89.97	94.70	31.70	56.48	73.47
BackTrans	75.87	76.21	79.20	44.85	50.86	54.19	70.46	72.76	78.93	89.86	92.34	94.36	53.56	64.52	73.13
EDA	75.49	77.64	79.14	47.13	50.15	53.39	77.66	84.46	87.37	90.20	92.11	94.93	47.00	64.15	73.53
AEDA	77.65	76.88	80.31	45.61	51.52	54.22	76.56	74.75	80.92	89.07	91.89	96.70	51.04	66.81	75.15
AMR-DA	77.49	76.32	77.93	48.80	52.37	54.68	77.98	78.37	86.54	93.69	94.03	96.28	52.82	64.02	72.09
SSMBA	76.64	77.4	79.85	46.95	50.53	53.97	78.64	83.92	85.94	90.31	89.75	93.69	47.07	60.99	70.24
GENIUS	77.52	77.71	78.35	51.90	51.69	51.46	77.32	75.72	78.64	93.58	94.14	96.70	51.76	65.34	73.17
PromDA	77.83	77.90	77.65	52.61	52.13	53.40	78.19	78.63	83.69	93.49	92.76	95.11	51.68	65.71	74.98
PromptMix	-	-	-	-	-	-	-	-	-	92.68	94.25	94.81	52.60	64.53	74.26
ZeroGen	73.84	75.66	76.30	41.47	49.21	54.55	68.42	80.19	86.79	81.24	83.95	85.63	28.20	47.02	67.80
GPT3Mix	57.87	61.80	66.12	31.60	32.98	50.33	62.58	74.90	80.73	76.91	81.75	85.36	25.91	46.72	68.99
CoDa *(ours)*	**79.70**	**80.11**	**81.20**	**53.70**	**54.32**	**55.81**	**84.58**	**86.72**	88.63	**93.92**	**94.45**	**96.82**	**54.64**	**67.74**	**76.20**
	±0.31	±0.26	±0.11	±0.52	±0.22	±0.31	±0.10	±0.69	±0.45	±0.18	±0.13	±0.04	±0.28	±0.15	±0.82

Table 2: Result comparison for Sequence Classification tasks. CoDa outperforms baselines by 0.12% - 5.94%.

Model	CoNLL-2003			OntoNotes			EBMNLP			BC2GM		
	100	200	500	100	200	500	100	200	500	100	200	500
Gold	52.89	66.53	70.43	16.37	27.7	61.46	14.83	21.3	27.8	47.46	54.38	59.41
LwTR	65.48	73.24	81.45	46.18	51.47	54.87	21.59	26.25	30.56	46.93	54.29	59.76
DAGA	53.91	51.63	54.68	33.29	43.07	54.64	10.97	14.89	18.90	34.67	41.98	48.72
MELM	56.89	62.23	79.05	11.94	31.55	45.68	18.29	22.01	25.12	40.86	51.32	55.79
GENIUS	67.85	58.2	80.36	25.08	23.29	22.14	20.08	16.87	21.41	43.41	52.01	56.65
CoDa *(ours)*	**70.45**	**80.43**	**84.23**	**48.19**	**53.81**	**62.78**	**23.22**	**27.12**	**32.45**	**49.56**	**54.85**	**61.11**
	±0.91	±0.84	±0.91	±0.45	±0.65	±0.72	±0.49	±0.79	±0.34	±0.54	±0.12	±0.42

Table 3: Result comparison for NER. CoDa outperforms baselines by 0.47% - 7.19%.

Model	SQuAD			NewsQA		
	100	200	500	100	200	500
Gold	11.64	19.71	26.32	22.45	30.14	45.65
BackTrans	17.47	22.60	29.07	27.32	34.98	47.21
EDA	17.07	22.39	28.98	29.31	35.81	49.90
AEDA	17.95	23.50	29.20	29.87	36.80	50.24
SSMBA	16.97	22.27	28.51	28.89	33.27	47.56
GENIUS	33.15	42.65	56.52	38.88	47.36	57.32
CoDa *(ours)*	**36.21**	**44.89**	**57.90**	**39.98**	**49.86**	**58.94**
	±0.21	±0.34	±0.11	±0.35	±0.15	±0.22

Table 4: Result comparison for QA. CoDa outperforms baselines by 1.10% - 3.06%.

Datasets. To demonstrate CoDa's flexibility, we evaluate it across various challenging datasets belonging to a wide range of domains. For SC, we employ Huffpost (Misra and Grover, 2021) (news category classification), Yahoo (Zhang et al., 2015) (answer topic classification), OTS (Drawzeski et al., 2021) (*legal* online service unfairness level classification), ATIS (Coucke et al., 2018) and Massive (FitzGerald et al., 2022) (Intent Classification). For NER, we employ ConLL-2003 (Tjong Kim Sang and De Meulder, 2003), OntoNotes-5.0 (Pradhan et al., 2013) (news domain), EBMNLP (Nye et al., 2018) and BC2GM (Krallinger et al., 2015) (bio-medical). Finally, for QA, we employ SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017). Details on each dataset and dataset statistics are provided in Section C.

Hyper-parameter settings. We prompt LLama-13B with a temperature of 0.5, top-p of 1.0, top-k=50. For all downstream NLU tasks, we employ $\text{BERT}_{\text{base-uncased}}$ (Devlin et al., 2019) as our encoder (except OTS where we employ legal-$\text{longformer}_{\text{large}}$ (Chalkidis* et al., 2023)). We fine-tuned our encoder with a batch size of 4,8 for 100 and 200 splits and 16 for 500 and 1000 splits. For NER specifically, we employ the flair library (Akbik et al., 2019) with an initial lr of $1e^{-5}$ and constant decay. Appendix A provides hyper-parameter tuning experiments. We report the micro-average F_1 score averaged across 3 runs for 3 random seeds.

5 Results and Analysis

Quantitative Analysis. Table 2, 3, and 4 compared CoDa with all our baselines on the tasks of SC, NER, and QA, respectively. CoDa outperforms our baselines in SC by 0.12% - 5.94%, NER by 0.47% - 7.19%, and QA by 1.10% - 3.06%. Though most prior methods proposed for one domain generally underperform in the other (Ghosh et al., 2023a), CoDa consistently outperforms these methods in all domains with varying semantic and syntactic properties, emphasizing its domain-agnostic nature.

Qualitative Analysis. Table 5 compares the generation quality of CoDa with all our baselines (averaged baseline-wise across all tasks and splits) on the measures of perplexity (Jelinek et al., 1977), diversity (average number of new tokens introduced in R augmentations) and length diversity (average absolute difference in length of source and R augmentations). CoDa outperforms most of our baselines in all settings. Additionally, as observed in Table 5, unlike other learning-based methods in

Original	If you are a consumer in the EEA, you can sue us in your home courts.	Coherent	Context Diversity	Label Consistency
Backtrans	If you are a consumer in the EEA, we do not exclude liability for gross negligence	✓	✗	✓
EDA	if you are a consumer in the eea you can sue us in your internal courts	✓	✗	✓
SSMBA	if you are a defendant in the eea, you can sue us in your own courts.	✓	✗	✓
AEDA	If you are a consumer . in ; the EEA, you can sue : us in your ! home courts.	✗	✗	✓
AMR-DA	If you are that consumer in the EEA , you can sue us in your own home court .	✓	✗	✓
GPT3Mix	Any consumer in EEA, can sue us in their home court.	✓	✗	✓
GENIUS	We are in the EEA, but we have to go to the courts in the UK.	✓	✗	✓
PromDA	EEA teams often find their greatest strengths playing on their home courts.	✓	✗	✓
CoDa	If you're a consumer and you sue us, these terms are fair. We'll only limit or exclude liability for death or personal injury caused by our negligence or fraud. You have statutory rights, and if you live in the EEA, Switzerland, Brazil, Hong Kong, Mexico, or Russia, different rules apply.	✓	✓	✓

Figure 2: Augmentation examples on the OTS dataset. All generations are produced in a low-resource setting (500 training examples). CoDa generates augmentations that are coherent, diverse, and label-consistent.

Method	Perplexity(↓)	Diversity(↑)	Perplexity(↓)	Diversity(↑)
	100		500	
EDA	104.93	115.89	118.83	156.21
GENIUS	24.90	120.64	25.43	126.32
GPT3Mix	88.77	146.89	75.17	163.32
BackTrans	240.93	132.51	74.91	56.31
AMR-DA	61.59	77.94	50.73	84.81
LwTR	135.89	94.77	139.93	99.63
CoDa *(ours)*	**22.44**	**152.34**	**23.33**	**165.81**

Table 5: Quantitative evaluation of generation quality on the measures of perplexity and token diversity. CoDa outperforms all our baselines on all metrics.

literature, the diversity of augmentations by CoDa does not depend on the number of gold training samples available. It performs equally well in both 100 and 500 splits.

Fig. 2 compares CoDa augmentations with other baselines in literature with a gold training sample taken from the OTS dataset. Generating augmentation on the OTS dataset, which belongs to the legal domain, is inherently difficult due to the formalized nature of legal language (Ghosh et al., 2023a). As we can see, CoDa generates augmentations that are coherent, diverse, and label-consistent. More examples are provided in Fig. 3, 4 and 5. Additionally, Appendix B evaluates how faithful LLaMa-2 was in following the constraints in the instructions.

6 Conclusion

We present CoDa, a simple and controllable data augmentation technique for low-resource NLP. CoDa extracts simple heuristic-based constraints from source sentences and verbalizes them to construct and instruction, which is then used to prompt LLMs to generate augmentations. CoDa is *training-free* and works with any out-of-the-box instruction-tuned LLM. Beyond providing explicit control, CoDa is also flexible, i.e., constraints can be easily replaced or added, enhancing its suitability across diverse domains.

Limitations and Future Work

Despite its effectiveness, CoDa suffers from various limitations, which we would like to mention. These limitations will remain our primary focus in future work. The limitations are as follows:

- LLMs often struggle to follow complex constraints in the instruction for text generation (Lu et al., 2023). We overcome this problem in CoDa by employing simple constraints. However, we acknowledge that data augmentation for complex domains and tasks may need to employ more complex constraints. Thus, as part of future work, we would like to employ recent advances in compositional prompting for breaking down complex constraints into simpler instructions.
- Although being *training free*, CoDa is computationally more expensive during inference time compared to prior art as it employs LLMs. As also shown in Section A.2, the overall performance of CoDa takes a slight hit when LLaMa-7B was employed instead of the 13B version. However, we acknowledge that as smaller models get better at following instructions, CoDa can perform more efficiently.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019.

FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Jiong Cai, Shen Huang, Yong Jiang, Zeqi Tan, Pengjun Xie, and Kewei Tu. 2023. Graph propagation based data augmentation for named entity recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 110–118, Toronto, Canada. Association for Computational Linguistics.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An empirical survey of data augmentation for limited data learning in nlp. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Shuguang Chen, Leonardo Neves, and Thamar Solorio. 2022. Style transfer as data augmentation: A case study on named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1827–1841, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Xiang Dai and Heike Adel. 2020. An analysis of simple data augmentation for named entity recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3861–3867, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 1–8, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Dan Friedman, Alexander Wettig, and Danqi Chen. 2022. Finding dataset shortcuts with grammar induction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4345–4363, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jonas Geiping, Micah Goldblum, Gowthami Somepalli, Ravid Shwartz-Ziv, Tom Goldstein, and Andrew Gordon Wilson. 2023. How much data are augmentations worth? an investigation into scaling laws, invariance, and implicit regularization. In *The Eleventh International Conference on Learning Representations*.

Sreyan Ghosh, Chandra Kiran Evuru, Sonal Kumar, S Ramaneswaran, S Sakshi, Utkarsh Tyagi, and Dinesh Manocha. 2023a. Dale: Generative data augmentation for low-resource legal nlp. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Sentosa, Singapore.

Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023b. Bioaug: Conditional generation based data augmentation for low-resource biomedical ner. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 1853–1858, New York, NY, USA. Association for Computing Machinery.

Sreyan Ghosh, Utkarsh Tyagi, Manan Suri, Sonal Kumar, S Ramaneswaran, and Dinesh Manocha. 2023c. Aclm: A selective-denoising based generative data augmentation approach for low-resource complex ner. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics.

Biyang Guo, Yeyun Gong, Yelong Shen, Songqiao Han, Hailiang Huang, Nan Duan, and Weizhu Chen. 2022. Genius: Sketch-based language model pre-training via extreme and selective masking for text generation and augmentation. *arXiv preprint arXiv:2211.10330*.

Xuming Hu, Yong Jiang, Aiwei Liu, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, and Philip S. Yu.

2023. Entity-to-text based data augmentation for various named entity recognition tasks.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(S1):S63–S63.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints.

Microsoft. 2023. Cntk: Language understanding/atis/data. Available at: `https://github.com/Microsoft/CNTK/tree/master/Examples/LanguageUnderstanding/ATIS/Data`.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1268–1283, Online. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Adir Rahamim, Guy Uziel, Esther Goldbraich, and Ateret Anaby Tavor. 2023. Text augmentation using dataset reconstruction for low-resource classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7389–7402, Toronto, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks. In *International Conference on Learning Representations*.

Gaurav Sahu, Olga Vechtomova, Dzmitry Bahdanau, and Issam H Laradji. 2023. Promptmix: A class boundary augmentation method for large language model distillation. *arXiv preprint arXiv:2310.14192*.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

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.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. PromDA: Prompt-based data augmentation for low-resource NLU tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4242–4255, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Honghua Zhang, Meihua Dang, Nanyun Peng, and Guy Van den Broeck. 2023. Tractable control for autoregressive language generation. In *International Conference on Machine Learning*, pages 40932–40945. PMLR.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang. 2022. FlipDA: Effective and robust data augmentation for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8646–8665, Dublin, Ireland. Association for Computational Linguistics.

Ran Zhou, Xin Li, Ruidan He, Lidong Bing, Erik Cambria, Luo Si, and Chunyan Miao. 2021. Melm: Data augmentation with masked entity language modeling for low-resource ner. *arXiv preprint arXiv:2108.13655*.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.

A Hyper-parameter Tuning

A.1 Effect of augmentation rounds R

Table 6 compares the performance of CoDa at different values of R. Augmenting the training dataset with several augmentation rounds R proves effective until the model overfits to the training data. The observation is similar to prior work in data augmentation for NLU tasks (Zhou et al., 2021; Ghosh et al., 2023c).

R	**1**	**2**	**3**	**4**	**5**	**6**	**7**
F_1	61.74	62.05	62.31	63.01	**63.16**	62.99	61.23

Table 6: F1 for various settings of R. All values are averaged across all datasets for all low-resource settings.

A.2 Choice of LLM

Table 7 compares the performance of CoDa employing different open-source LLMs. Beyond LLaMa-13B employed in our paper, we also compare performance with Mistral-7B (Jiang et al., 2023) and LLaMa-7B. As we see, employing LLaMa-7B takes a hit of 0.18% on the final performance, while employing Mistral-7B takes a hit of 1.44% on the final performance. We also noticed several instances of hallucination with Mistral-7B, where the output of the LLM was completely different from the given instruction. This was not the case with the LLaMa family of models, and performance generally improved with a larger model owing to a better quality of generations and better abilities to follow instructions.

LLM	**F1-Micro**
Mistral-7B	61.72
LLaMa-7B	62.98
LLaMa-13B	**63.16**

Table 7: F1 micro averaged across tasks for various LLMs.

B Faithfulness in following instruction constraints

Table 8 illustrates the accuracy of augmentations produced by LLaMa-13B in adhering to the constraints specified in the instruction. We only illustrate accuracies of Lexical and Length constraints as they are easily quantifiable. Other constraints require human evaluation, which remains part of future work. We report the accuracy as our metric for faithfulness, wherein we consider a generation as accurate for the constraint if it completely follows the constraint, else inaccurate. Additionally, we also report a 75% threshold for both the constraints, whereby we consider the generation as accurate if it follows 75% of the constraint (e.g.,75% of the total keywords mentioned are in generation and the total tokens in the generation lie between 75% of the maximum and minimum lengths). Although LLaMa-13B demonstrates moderate proficiency in adhering to constraints, the anticipated improvement in instruction-following capabilities of LLMs is likely to enhance these metrics further. Furthermore, the fact that CoDa surpasses the performance of many existing models in the literature, despite its moderate ability to follow constraints, suggests a significant promise for CoDa as an augmentation generation scheme when integrated with more advanced LLMs.

An observed trend is that models demonstrate strong performance on familiar datasets such as CoNLL-2003, potentially due to these datasets being included in their pre-training corpus. Additionally, our models exhibit improved performance under 75% threshold constraints. This suggests a balance must be struck between the creative output and adherence to constraints in LLM generations. Although creativity is crucial for generating diverse augmentations, following constraints is key for maintaining consistency. In future work, we aim to investigate more effective methods for balancing this trade-off.

Task	Lexical	Lexical 75%	Length	Length 75%
HuffPost	24.64	26.09	51.31	55.02
Yahoo	27.28	28.12	51.06	54.48
OTS	21.83	23.32	50.98	53.95
ATIS	41.1	43.5	50.2	51.52
MASSIVE	26.26	28.32	50.22	51.52
CoNLL-2003	67.72	73.31	51.13	53.82
OntoNotes	36.33	48.7	50.59	53.12
EBMNLP	41.05	45.46	50.72	53.17
BC2GM	41.45	48.82	50.6	53.17
SQUAD	32.56	40.87	52.12	55.82
NEWSQA	33.45	42.18	51.98	54.87

Table 8: Faithfulness of generated augmentations. Scores reported correspond to average accuracy, where we attribute an augmentation as accurate if it perfectly follows the constraint in the given instruction; otherwise, we attribute it as inaccurate.

C Dataset Details

C.1 Classification

HuffPost. The HuffPost dataset (Misra and Grover, 2021) is a popular multiclass classification dataset in NLP. It is a collection of news articles from the HuffPost website, covering a wide range of topics, including politics, business, entertainment, and more. For multiclass classification, the HuffPost dataset is labeled with a diverse set of categories and for our experiments, we take sentences from five categories, including politics, sports, entertainment, tech, and business. Dataset statistics can be found in Table 9.

Yahoo. The Yahoo Answers topic classification dataset (Zhang et al., 2015) is a widely used dataset for multi-class text classification tasks. It is derived from the Yahoo Answers community-driven question-answering platform, where users ask questions on various topics, and community members provide answers. The dataset contains a large number of question-and-answer pairs covering a wide range of categories or topics. Each question in the dataset is associated with one primary category. The primary categories span diverse subjects, including Society & Culture, Science & Mathematics, Health, Education & Reference, Computers & Internet, Sports, Business & Finance, Entertainment & Music, Family & Relationships, Politics & Government, Travel, Cars & Transportation, Food & Drink, Games & Recreation, Home & Garden, Local Businesses, News & Events, Pets, Beauty & Style and Pregnancy & Parenting. Dataset statistics can be found in Table 9.

OTS-UL. Online Terms of Service (OTS) (Drawzeski et al., 2021) attempt to automatically detect unfair clauses in Terms of Service. The input to the model is a sentence, and the output presents the sentence classified into three levels of unfairness. The dataset setting used in our paper is similar to (Niklaus et al., 2023). Dataset statistics can be found in Table 9.

C.2 Named Entity Recognition

CoNLL-2003. The CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003) is a widely used benchmark dataset for Named Entity Recognition

Dataset	Source	Sub-domain	Task Type	Training/Dev/Test Instances	Classes
HuffPost	Misra and Grover (2021)	HuffPost website	Multi-class classification	67490/16891/16891	5
Yahoo	Zhang et al. (2015)	Yahoo Answers	Multi-class classification	1375404/58966/58966	10
OTS-UL	Drawzeski et al. (2021)	EU Law	Multi-class classification	2074/191/417	3
ATIS	Microsoft (2023)	Travel enquiry	Intent Classification	4972/888/888	17
MASSIVE	FitzGerald et al. (2022)	Diverse	Intent Classification	11500/2030/2970	60
CoNLL-2003	Tjong Kim Sang and De Meulder (2003)	English news articles	Named Entity Recognition	14041/3250/3453	4
OntoNotes-5.0	Pradhan et al. (2013)	Diverse	Named Entity Recognition	115812/15680/12217	36
BC2GM	Krallinger et al. (2015)	Biomedical	Named Entity Recognition	15197/3061/6325	2
EBMNLP	Nye et al. (2018)	Biomedical	Named Entity Recognition	35005/10123/6193	7
SQUAD	(Rajpurkar et al., 2016)	Wikipedia Articles	Question Answering	87600/10600/-	-
NEWSQA	(Trischler et al., 2017)	CNN Articles	Question Answering	92549/5126/5166	-

Table 9: Statistics for each downstream NLU datasets used in our experiments. As described in Section 4, we derive low-resource splits from these original datasets for our experiments.

(NER) tasks in NLP. It was created for the Conference on Computational Natural Language Learning (CoNLL) shared task in 2003. The dataset consists of news articles from the Reuters Corpus, a collection of English news articles. It is annotated with four named entities: person, organization, location, and miscellaneous entities (such as dates and percentages). The annotations indicate the boundaries of the named entities within the text. Dataset statistics can be found in Table 9.

Ontonotes 5.0. Ontonotes 5.0 Pradhan et al. (2013) is a widely used dataset in the field of Natural Language Processing (NLP) and specifically for Named Entity Recognition (NER) tasks. It is a large-scale corpus that provides annotations for a variety of linguistic phenomena, including named entities, across multiple languages. The dataset contains a diverse range of text genres, including news articles, conversational data, and web data, making it suitable for training and evaluating NER models in different domains. It covers three languages: English, Chinese, and Arabic. The dataset is annotated with 11 categories: Person, Organization, Location, Date, Time, Money, Percent, Quantity, Ordinal and Miscellaneous. Dataset statistics can be found in Table 9.

EBMNLP. EBMNLP Nye et al. (2018) is a widely used dataset in the field of Biomedical Named Entity Recognition (BioNER) tasks. It is a corpus of richly expert-annotated abstracts of medical articles describing clinical randomized controlled trials. The dataset facilitates easy search and organization of published literature on randomized controlled trials, addressing the current challenges impeding the goals of evidence-based medicine (EBM). The dataset is annotated with 3 categories: Outcome, Intervention and Participant. Dataset statistics can be found in Table 9.

BC2GM. BC2GM Krallinger et al. (2015) is a widely used dataset in the field of Biomedical Named Entity Recognition (BioNER) tasks. This dataset is a part of the CHEMDNER large scale corpus which includes annotation of chemical entities as well as named entities in the biomedical and other domains. The dataset is annotated with 1 categoriy: Gene. Dataset statistics can be found in Table 9.

C.3 Intent Classification

ATIS. The ATIS (Airline Travel Information System) dataset[2] is a widely used benchmark dataset for intent classification in the field of NLU. It was developed to address understanding user intents in the context of airline travel information. The dataset consists of queries or utterances that users might input when interacting with a flight reservation system. Each query is labeled with an intent representing the user's intention or purpose behind the query. The dataset is labeled with intents that are: Flight-Booking, Flight-Status, Flight-Information, Ground-Service, Airfare, Airport-Information, Travel-Preferences, Flight-Cancellation, and None/No-Intent. Dataset statistics can be found in Table 9.

MASSIVE. The MASSIVE (Multilingual Amazon Slu resource package for Slot-filling) FitzGerald et al. (2022) dataset is a widely used benchmark dataset for intent classification in the field of NLU. It contains 1M realistic, parallel, labeled virtual assistant utterances spanning 51 languages, 18 domains, 60 intents, and 55 slots. The dataset is labeled with intents some of which are: Alarm set, Play music, Audio volume mute, Weather query, Takeaway order and General joke etc. Dataset statistics can be found in Table 9.

C.4 Question Answering

SQUAD. The SQUAD (Stanford Question Answering Dataset) (Rajpurkar et al., 2016) is a read-

[2] https://github.com/howl-anderson/ATIS_dataset/tree/master

ing comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. Dataset statistics can be found in Table 9.

NEWSQA. NewsQA (News Question Answering) (Trischler et al., 2017) is a challenging machine comprehension dataset of over 100,000 human-generated question-answer pairs. Crowdworkers supply questions and answers based on a set of over 10,000 news articles from CNN, with answers consisting of spans of text from the corresponding articles. Dataset statistics can be found in Table 9.

D Baseline Details

SSMBA. SSMBA (Ng et al., 2020) generates synthetic training examples by using a pair of corruption and reconstruction functions to move randomly on a data manifold.

AEDA. AEDA (Karimi et al., 2021) is similar to EDA but only employs random insertion of punctuation marks in the original text to generate synthetic augmentations.

GENIUS. GENIUS (Guo et al., 2022), pre-trains and optionally fine-tunes BART (Lewis et al., 2019) on a denoising objective using sketches generated with an extreme masking algorithm. The extreme masking algorithm just preserves keywords in a sentence and masks everything else.

MELM. MELM (Zhou et al., 2021), which stands for Masked Entity Language Modeling, suggests the fine-tuning of a transformer-encoder-based PLM on linearized labeled sequences through masked language modeling. In low-resource scenarios, MELM surpasses all other baselines and prior techniques on the CoNLL 2003 NER dataset across four languages, including mono-lingual, cross-lingual, and multi-lingual settings.

DAGA. DAGA (Ding et al., 2020), short for Data Augmentation with a Generation Approach, suggests the training of a one-layer LSTM-based recurrent neural network language model (RNNLM) by maximizing the probability of predicting the next token using linearized sentences. For sentence generation, they employ random sampling to create entirely new sentences, with the model being fed only the [**BOS**] token.

LwTR. LwTR (Dai and Adel, 2020) replaces a token in a sentence with another token of the same label; the token is randomly selected from the training set.

PromDA. PromDA (Wang et al., 2022) proposes a data augmentation framework based on T5 that trains soft prompts using a novel keyword-to-sentence algorithm.

AMR-DA. AMR-DA (Shou et al., 2022) converts a sample document from a dataset to an AMR graph, modifies the graph according to various data augmentation policies, and then generates augmentations from graphs. The method combines both sentence-level techniques like back translation and token-level techniques like EDA.

PromptMix. PromptMix (Sahu et al., 2023) PromptMix prompts instruction-tuned LLMs to generate augmentations for text classification tasks that are close to the class boundary.

ZeroGen. ZeroGen (Ye et al., 2022), similar to PromptMix, generates data using LLMs but in a zero-shot manner without any gold data. It prompts pre-trained LLMs (not instruction fine-tuned) for data synthesis.

We do not consider more recent baselines provided by Cai et al. (2023), Hu et al. (2023) and Rahamim et al. (2023) as the code for the same was not available at the time of writing the paper. Additionally, we do not consider Zhou et al. (2022) as label flipping is not applicable for our paper for all tasks considered, and Chen et al. (2022) as style transfer is better suited for cross-domain tasks and applying it to single domain tasks is not trivial. Finally, we do not consider Yu et al. (2023) as it requires manual human intervention for attribute extraction for a dataset.

E Additional Details

E.1 Examples of syntactic constraints in formal domains

Table 10 provides examples of documents from domains with formal language, like legal and biomedical. Each example provides two corresponding documents to a POS sequence, emphasizing that syntactic constraints help generate augmentations better aligned to the domain in formal domains.

E.2 Qualitative Examples

Fig. 2, 3, 4 and 5 provide additional qualitative examples of augmentations generated using CoDa and compares them with other baselines. CoDa

Dataset	Syntactic Examples
OTS	1. **Constraint**: ADV PRON AUX PUNCT ADP PRON NOUN PUNCT VERB NOUN ADP DET NOUN PUNCT **Generation 1**: Quickly he can, upon her request, examine the document. **Generation 2**: Additionally, he shall submit the document to the court. 2. **Constraint**: NOUN AUX PART ADJ ADP NOUN PRON AUX PART VERB PRON NOUN PUNCT **Generation 1**: Contractors have, under new regulations, completed their work. **Generation 2**: Judge may have been impartial in the legal proceedings.
EBMNLP	1. **Constraint**: DET ADJ NOUN AUX VERB ADP NOUN NOUN PUNCT **Generation 1**: The molecular analysis revealed a genetic mutation in the patient. **Generation 2**: The experimental procedure was conducted on laboratory samples. 2. **Constraint**: NOUN NOUN NOUN ADP NUM NOUN AUX VERB VERB ADP NUM NOUN NOUN PUNCT **Generation 1**: EXPERIMENT A cohort of 50 samples was collected from 3 laboratory facilities. **Generation 2**: STUDY A group of 100 patients underwent testing in two medical centers.
BC2GM	1. **Constraint**: PROPN PROPN PROPN ADP DET NOUN ADP NOUN NOUN ADP NOUN PUNCT **Generation 1**: Polymerase chain reaction for the detection of genetic mutations in patients. **Generation 2**: Hormone receptor status in the evaluation of breast cancer in women. 2. **Constraint**: NOUN ADP NOUN NOUN CCONJ NOUN PUNCT **Generation 1**: Analysis of protein structures and 3,4-dihydroxyphenylalanine. **Generation 2**: Exploration of biochemical pathways and 2,3-dimethylbutane.

Table 10: Examples of a couple of documents corresponding to a single POS sequence in formal domains like legal (OTS) and bio-medical (EBMNLP and BC2GM). We emphasize that syntactic constraints help generate augmentations better aligned to the domain.

consistently generates more diverse and consistent augmentations over prior art.

F Extra Details

F.1 Model Parameters

$\text{BERT}_{\text{base}}$ has ≈ 110M 12-layers of encoder, 768-hidden-state, 2048 feed-forward hidden-state, and 8-heads. legal-$\text{longformer}_{\text{large}}$ has ≈ 149M 30 layers of encoder, 768-hidden-state, 3072 feed-forward hidden-state, and 12-heads. LLaMa-13B is a 13B parameter model and LLaMa-7B is a 7B parameter model.

F.2 Compute Infrastructure

All our experiments are conducted on NVIDIA A100 and NVIDIA A6000 GPUs. We batch prompted LLaMa-2 13B and LLaMa-2 7B, with a BS of 16, where LLaMa-2 performed distributed inference on 4 A6000 GPUs. Fine-tuning on the downstream tasks uses 4 A100 GPUs.

F.3 Implementation Software and Packages

We implement all our models in PyTorch[3] and use the HuggingFace[4] implementations of $\text{BERT}_{\text{base}}$, legal-$\text{longformer}_{\text{large}}$, LLaMa-13B and LLaMa-7B. For NER specifically, we employ the Flair [5] library.

We also use the following repositories for running the baselines: BackTrans (Yu et al., 2018), EDA[6](Wei and Zou, 2019), AEDA[7] (Karimi et al., 2021), AMR-DA[8] (Shou et al., 2022), SSMBA[9] (Ng et al., 2020), GENIUS(-**ft**)[10] (Guo et al., 2022), PromDA[11] (Wang et al., 2022), PromptMix[12] (Sahu et al., 2023), ZeroGen[13] (Ye et al., 2022), GPT3Mix[14] (Yoo et al., 2021), LwTR[15] (Dai and Adel, 2020), DAGA[16] (Ding et al., 2020)(Ding et al., 2020) and MELM[17] (Zhou et al., 2021). All the baseline repositories are covered under the MIT License.

F.4 Dataset Links

We use the following datasets to evaluate: Huffpost[18] (Misra and Grover, 2021), Yahoo[19] (Zhang et al., 2015), OTS[20] (Drawzeski et al., 2021), Massive[21] (FitzGerald et al., 2022), ATIS[22] (Coucke

[3] https://pytorch.org/
[4] https://huggingface.co/
[5] https://github.com/flairNLP/flair

[6] https://github.com/jasonwei20/eda_nlp
[7] https://github.com/akkarimi/aeda_nlp
[8] https://github.com/zzshou/amr-data-augmentation
[9] https://github.com/nng555/ssmba
[10] https://github.com/beyondguo/genius
[11] https://github.com/GaryYufei/PromDA
[12] https://github.com/servicenow/promptmix-emnlp-2023
[13] https://github.com/jiacheng-ye/ZeroGen
[14] https://github.com/naver-ai/hypermix
[15] https://github.com/boschresearch/data-augmentation-coling2020
[16] https://github.com/ntunlp/daga
[17] https://github.com/randyzhouran/melm
[18] https://www.kaggle.com/datasets/rmisra/news-category-dataset
[19] https://huggingface.co/datasets/yahoo_answers_topics
[20] https://huggingface.co/datasets/joelniklaus/lextreme
[21] https://huggingface.co/datasets/AmazonScience/massive
[22] https://github.com/howl-anderson/ATIS_dataset

et al., 2018), ConLL-2003[23] (Tjong Kim Sang and De Meulder, 2003), OntoNotes-5.0[24] (Pradhan et al., 2013), EBMNLP[25] (Nye et al., 2018) and BC2GM[26] (Krallinger et al., 2015), SQuAD[27] (Rajpurkar et al., 2016) and NewsQA[28] (Trischler et al., 2017). All the datasets have been released under various licenses for research purposes.

F.5 Potential Risks

Diffusion models learn from vast amounts of textual data, including biased or prejudiced content present on the internet. As a result, there is a risk of bias amplification, where the models unintentionally perpetuate or reinforce existing biases. Also, diffusion models can generate highly coherent and contextually plausible text, raising concerns regarding the potential for generating misinformation or disinformation.

[23] https://huggingface.co/datasets/conll2003

[24] https://catalog.ldc.upenn.edu/LDC2013T19

[25] https://huggingface.co/datasets/bigbio/ebm_pico

[26] https://huggingface.co/datasets/bc2gm_corpus

[27] https://rajpurkar.github.io/SQuAD-explorer

[28] https://www.microsoft.com/en-us/research/project/newsqa-dataset/download/

Original	list all the takeoffs and landings at general mitchell international	Coherent	Context Diversity	Label Consistency
BackTrans	List of all take	✗	✗	✓
EDA	list all the takeoffs and landings at general mitchell astatine international	✓	✗	✓
SSMBA	list all the compilers and events at general, international	✓	✗	✗
AEDA	list all the : takeoffs and landings at general? mitchell international !	✗	✗	✓
AMR-DA	What is the general Mitchell International list of all takeoffs and landings ?	✓	✗	✓
GPT3Mix	list all the timings for arrival and departures at general mitchell international	✓	✗	✓
GENIUS	All the news, all the takeoffs and landings. All the news you need to know.	✓	✗	✓
PromDA	there are many takeoffs and landings at general mitchell	✓	✗	✓
CoDa	Experience the thrill of flights from Atlanta to Boston, with convenient takeoffs and landings at General Mitchell International Airport	✓	✓	✓

Figure 3: Augmentation examples on the ATIS dataset. All generations are produced in a low-resource setting (500 training examples).

Original	Iraq's President Saddam Hussein meets with chairman of the Russian liberal democratic party Vladimir Zhirinovsky .	Coherent	Context Diversity	Label Consistency
LwTR	U.S. changes President Hussein meets with chairman of the Russian liberal democratic lift Vladimir Zhirinovsky evidence.	✓	✗	✓
DAGA	Saddam Hussein meets with chairman of the Russian party Vladimir Zhirinovsky .	✓	✗	✓
MELM	President has told visiting ultra-nationalist Zhiriddisky to talk about ties	✗	✗	✓
GENIUS	Iraqi President Saddam Hussein says he has met with a Russian, Vladimir Zhirinovsky, in Baghdad and Moscow.	✓	✗	✓
CoDa	Saddam Hussein, the Iraqi President, met with Vladimir Zhirinovsky, the leader of the Russian Liberal Democratic Party, in Baghdad to discuss strengthening ties between the two nations.	✓	✓	✓

Figure 4: Augmentation examples on the CoNLL-2003 dataset. All generations are produced in a low-resource setting (500 training examples).

Original	Document: Beyoncé's first solo album Dangerously in Love was released on June 24, 2003, after Michelle Williams and Kelly Rowland had released their solo efforts. The album sold 317,000 copies in its first week, debuted atop the Billboard 200, and has since sold 11 million copies worldwide. The album's lead single, "Crazy in Love", featuring Jay Z, became Beyoncé's first number-one single as a solo artist in the US. The single "Baby Boy" also reached number one. Question: What solo album did Beyonce release in 2003?	Coherent	Context Diversity	Label Consistency
Backtrans	Question: What solo album was released by Beyonce in 2003?	✓	✗	✓
EDA	Question: What single album was published by Beyonce in 2003?	✓	✗	✓
SSMBA	Question: In 2003, which solo album did Beyoncé release?	✓	✗	✓
AEDA	Question: *What solo . album! did Beyonce release in .2003?	✗	✗	✓
GENIUS	Document: "Beyonce first solo recording was released in October 2002. Her first solo album Dangerously in Love was released on June 24, 2003, after Michelle Williams and Kelly Rowland had released their solo efforts. The followup became Beyonces most successful single to date, selling more than 10 million copies worldwide.	✓	✗	✓
CoDa	Document: Beyoncé released her first solo album, "Dangerously in Love," on June 24, 2003. The album was a commercial success, earning her a Grammy and selling 317,000 copies in its first week. The album features a mix of Rap/Sung and Vocals, and includes the hit singles "Crazy in Love" and "Baby Boy."	✓	✓	✓

Figure 5: Augmentation examples on the SQUAD dataset. All generations are produced in a low-resource setting (500 training examples).

Method	Sentence	Instruction 1	Instruction 2
Yahoo	Shops in most malls advertise for Christmas help up to the last minute.	Write a brief document with a single sentence or multiple sentences with the following constraints: 1. The document should have the following keywords: advertise or marketing, Shops, malls. 2. The document should be on the topic of Business & Finance. 3. The document should have a length of 13-19 words. 4. Any sentence in the document should not include the abstract concept coaching. 5. Any sentence in the document should not include the abstract concept market volatility. 6. Any sentence in the document should not include the abstract concept market share.	Write a brief document with a single sentence or multiple sentences corresponding to the following abstract description: "Christmas help wanted ads in malls often run until the last minute." . Additionally, the document should have the following constraints: 1. The document should have the following keywords: business, industry, marketing, profits, but should not have the following keywords : develop. 2. The document should be on the topic of Business & Finance. 3. The document should have a length of 413-619 words. 4. Any sentence in the document should not include the abstract concept coaching. 5. Any sentence in the document should not include the abstract concept market volatility. 6. Any sentence in the document should not include the abstract concept market share.
OTS	We are not obligated to publish any information or content on our Service and can remove it with or without notice.	Write a brief document with a single sentence or multiple sentences with the following constraints: 1. The document should have the following keywords: obligated,notice or prejudice,Service, but should not have the following keywords: responsible, liable. 2. The document's terms of service should be clearly unfair. 3. The document should have a part-of-speech sequence similar to: PRON AUX PART VERB PART VERB DET NOUN CCONJ NOUN ADP PRON PROPN CCONJ AUX VERB PRON ADP CCONJ ADP NOUN PUNCT. 4. The document should have a length of 21-31 words. 5. Any sentence in the document should not include the abstract concept litigation. 6. Any sentence in the document should not include the abstract concept account management. 7. Any sentence in the document should not include the abstract concept jurisdiction.	Write a brief document with a single sentence or multiple sentences corresponding to the following abstract description: "We reserve the right to remove content without notice.". Additionally, the document should have the following contraints: 1. The document should have the following keywords: reason,right,way or data, but should not have the following keywords: cause. 2. The document's terms of service should be clearly unfair. 3. The document should have a part-of-speech sequence similar to: PRON VERB DET NOUN PART VERB CCONJ VERB PROPN PROPN PUNCT CCONJ VERB DET NOUN PRON VERB ADP PROPN PUNCT ADP DET NOUN PUNCT. 4. The document should have a length of 21-31 words. 5. Any sentence in the document should not include the abstract concept litigation. 6. Any sentence in the document should not include the abstract concept account management. 7. Any sentence in the document should not include the abstract concept jurisdiction.
CoNLL-2003	Israel approves Arafat's flight to West Bank.	Write a brief document with a single sentence or multiple sentences with the following constraints: 1. The document should have the following keywords: Israel,Arafat,West Bank,approves or confirms. 2. Israel is location, Arafat is person, West Bank is location. 3. The document should have a length of 5-13 words.	
BC2GM	Comparison with alkaline phosphatases and 5 - nucleotidase	Write a brief document with a single sentence or multiple sentences with the following constraints: 1. The document should have the following keywords: alkaline phosphatases,5 - nucleotidase,Comparison. 2. alkaline phosphatases is a Gene. 3. The document should have a part-of-speech sequence similar to: NOUN ADP ADJ NOUN CCONJ NUM PUNCT NOUN. 4. The document should have a length of 5-12 words.	

Table 11: Instruction prompts for various tasks.

Method	Sentence	Instruction 1	Instruction 2
SQUAD	Beyoncé's first solo recording was a feature on Jay Z's "'03 Bonnie & Clyde" that was released in October 2002, peaking at number four on the U.S. Billboard Hot 100 chart. Her first solo album Dangerously in Love was released on June 24, 2003, after Michelle Williams and Kelly Rowland had released their solo efforts. The album sold 317,000 copies in its first week, debuted atop the Billboard 200, and has since sold 11 million copies worldwide. The album's lead single, "Crazy in Love", featuring Jay Z, became Beyoncé's first number-one single as a solo artist in the US. The single "Baby Boy" also reached number one, and singles, "Me, Myself and I" and "Naughty Girl", both reached the top-five. The album earned Beyoncé a then record-tying five awards at the 46th Annual Grammy Awards; Best Contemporary R&B Album, Best Female R&B Vocal Performance for "Dangerously in Love 2", Best R&B Song and Best Rap/Sung Collaboration for "Crazy in Love", and Best R&B Performance by a Duo or Group with Vocals for "The Closer I Get to You" with Luther Vandross.	Write a brief document with multiple sentences corresponding to the following constraints: 1. The document should have the following keywords 11,Vocals,Hot,copies,lead,Baby,also,Vandross, You,Album,Best,earned,Rap/Sung,Grammy, Clyde,"Her first solo album Dangerously in Love was released on June 24, 2003, after Michelle Williams and Kelly Rowland had released their solo efforts". 2. The document should have a length of 113-340 words.	

Table 12: Instruction prompts for SQUAD dataset.