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, utkarsh, dmanocha}@umd.edu

Abstract

We present CoDa (Constrained Generation based Data 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.

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

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

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 Pre-trained 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).

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Original 1: Scott Gimple sets the record on this big midseason finale scene</th>
<th>Original 2: mine is all of Isaiah chapter 6 because that is what Jesus did for me.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BackTrans</td>
<td>1. Scott Gimple sets the record on this big midseason finale scene</td>
<td>2. mine is all of Isaiah chapter 6 because that is what Jesus did for me.</td>
</tr>
<tr>
<td>(Tian et al., 2019)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDA</td>
<td>1. mine is all of Isaiah chapter 6 because that is what Jesus did for me.</td>
<td></td>
</tr>
<tr>
<td>(Wei and Zou, 2019)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSMRA</td>
<td>1. mine is all of Isaiah chapter 6 because that is what Jesus did for me.</td>
<td></td>
</tr>
<tr>
<td>(Ng et al., 2020)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPT3Mix</td>
<td>1. Scott Gimple sets the record on this big midseason finale scene</td>
<td></td>
</tr>
<tr>
<td>(Chen et al., 2021)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GENUS</td>
<td>1. For the record, there was no midseason finale scene. This is, until the finale</td>
<td></td>
</tr>
<tr>
<td>(Guo et al., 2022)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CoDa (ours)</td>
<td>1. The recording of the scene in the new movie by Scott Gimple was a big success, capturing the essence of the entertaining story.</td>
<td>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.</td>
</tr>
</tbody>
</table>
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 \textit{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, \textit{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., 2023a,b). Secondly, \textit{text editing} focuses on modifying certain parts of a given sentence (Wei and Zou, 2019; Shou et al., 2022). Thirdly, \textit{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 $D = \{d_1, \ldots, d_i, \ldots, d_n\}$, 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 $D$. For the latter, we first retrieve a document from $D$, 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 $D$, 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 $D$ 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) $N = \{n_0, \cdots, n_i, \cdots, n_T\}$. Next, we assign an importance score to each by calculating cosine similarity between $E(n_i)$ and $E(d)$, where $E$ 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 $D$ 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 $D$.

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 $D$ 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(Sc-f) (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.
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 (Krarliger 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 BERTbase-uncased (Devlin et al., 2019) as our encoder (except OTS where we employ legal-longformstf3 (Chalkidis et al., 2023)). We finetuned 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 (Akibik et al., 2019) with an initial lr of $10^{-5}$ 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...
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.

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.


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.


Vancouver, Canada. Association for Computational Linguistics.


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).

<table>
<thead>
<tr>
<th>$R$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_1$</td>
<td>61.74</td>
<td>62.05</td>
<td>62.31</td>
<td>63.01</td>
<td><strong>63.16</strong></td>
<td>62.99</td>
<td>61.23</td>
</tr>
</tbody>
</table>

Table 6: $F_1$ 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.

<table>
<thead>
<tr>
<th>LLM</th>
<th>F1-Micro</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mistral-7B</td>
<td>61.72</td>
</tr>
<tr>
<td>LLaMa-7B</td>
<td>62.98</td>
</tr>
<tr>
<td>LLaMa-13B</td>
<td><strong>63.16</strong></td>
</tr>
</tbody>
</table>

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.

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


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

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.
(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 category: Gene. Dataset statistics can be found in Table 9.

### C.3 Intent Classification

**ATIS.** The ATIS (Airline Travel Information System) dataset 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-

---

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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source</th>
<th>Sub-domain</th>
<th>Task Type</th>
<th>Training/Dev/Test Instances</th>
<th>Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>Zhang et al. (2015)</td>
<td>Yahoo Answers</td>
<td>Multi-class classification</td>
<td>1375404/58966/58966</td>
<td>10</td>
</tr>
<tr>
<td>OTS-UL</td>
<td>Drawzinski et al. (2021)</td>
<td>EU Law</td>
<td>Multi-class classification</td>
<td>2074/191/417</td>
<td>3</td>
</tr>
<tr>
<td>ATIS</td>
<td>Microsoft (2023)</td>
<td>Travel enquiry</td>
<td>Intent Classification</td>
<td>4972/888/888</td>
<td>17</td>
</tr>
<tr>
<td>MASSIVE</td>
<td>FitzGerald et al. (2022)</td>
<td>Diverse</td>
<td>Intent Classification</td>
<td>115900/2030/2970</td>
<td>60</td>
</tr>
<tr>
<td>OntoNotes-5.0</td>
<td>Pradhan et al. (2013)</td>
<td>Diverse</td>
<td>Named Entity Recognition</td>
<td>115812/15680/12217</td>
<td>36</td>
</tr>
<tr>
<td>EBMNLP</td>
<td>Nye et al. (2018)</td>
<td>Biomedical</td>
<td>Named Entity Recognition</td>
<td>35005/10123/6193</td>
<td>7</td>
</tr>
<tr>
<td>SQUAD</td>
<td>Rajpurkar et al., (2016)</td>
<td>Wikipedia Articles</td>
<td>Question Answering</td>
<td>87690/10680/-</td>
<td>-</td>
</tr>
<tr>
<td>NEWSQA</td>
<td>Trischler et al., (2017)</td>
<td>CNN Articles</td>
<td>Question Answering</td>
<td>92549/5126/5166</td>
<td>-</td>
</tr>
</tbody>
</table>

2https://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) 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
Table 10: Examples of a couple of documents corresponding to a single POS sequence in formal domains like legal (OTS) and bio-medical (EBMNLP). We emphasize that syntactic constraints help generate augmentations better aligned to the domain.

Download the following datasets to evaluate: Huff-
dataset2020 (Drawzeski et al., 2021), Mas-
eda-20 (Sahu et al., 2023), ZeroGen
ft(Wei and Zou, 2019), AEDA7 (Karimi
e et al., 2021), AMR-DA8 (Shou et al., 2022),
SSMBA9 (Ng et al., 2020), GENIUS(-ft)10 (Guo
e et al., 2022), PromDA11 (Wang et al., 2022),
PromptMix12 (Sahu et al., 2023), ZeroGen13 (Ye
et al., 2022), GPT3Mix14 (Yoo et al., 2021),
LwTR15 (Dai and Adel, 2020), DAGA16 (Ding
e et al., 2020)(Ding et al., 2020) and MELM17 (Zhou
et al., 2021). All the baseline repositories are cov-
ered under the MIT License.

F Extra Details

F.1 Model Parameters
BERTbase has = 110M 12-layers of encoder, 768-
hidden-state, 2048 feed-forward hidden-state, and
8-heads. legal-longformeralarge 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 PyTorch3 and use
the HuggingFace4 implementations of BERTbase,
legal-longformeralarge, LLaMa-13B and LLaMa-7B.
For NER specifically, we employ the Flair5 library.
We also use the following repositories for run-
nig the baselines: BackTrans (Yu et al., 2018),
EDA6(Wei and Zou, 2019), AEDA7 (Karimi
et al., 2021), AMR-DA8 (Shou et al., 2022),
SSMBA9 (Ng et al., 2020), GENIUS(-ft)10 (Guo
et al., 2022), PromDA11 (Wang et al., 2022),
PromptMix12 (Sahu et al., 2023), ZeroGen13 (Ye
et al., 2022), GPT3Mix14 (Yoo et al., 2021),
LwTR15 (Dai and Adel, 2020), DAGA16 (Ding
e et al., 2020)(Ding et al., 2020) and MELM17 (Zhou
et al., 2021). All the baseline repositories are cov-
ered under the MIT License.

F.4 Dataset Links
We use the following datasets to evaluate: Huff-
dataset2020 (Misra and Grover, 2021), Yahoo19 (Zhang
e et al., 2015), OTS20 (Drawzeski et al., 2021), Massive21 (FitzGerald et al., 2022), ATIS22 (Coucke

1https://github.com/jasonwei20/eda_nlp
2https://github.com/akkarimi/aeda_nlp
3https://github.com/zzshou/amr-data-augmentation
4https://github.com/nng555/ssmba
5https://github.com/beyondguo/genius
6https://github.com/GaryYufei/PromDA
7https://github.com/servicenow/promptmix-emnlp-2023
8https://github.com/jiacheng-ye/ZeroGen
9https://github.com/naver-ai/hypermix
11https://github.com/ntunlp/daga
12https://github.com/randyzhouran/melm
13https://www.kaggle.com/datasets/rmisra/news-category-dataset
14https://huggingface.co/datasets/yahoo_answers_topics
15https://huggingface.co/datasets/joelniklaus/extreme
16https://huggingface.co/datasets/AmazonScience/massive
17https://github.com/howl-anderson/ATIS_dataset
et al., 2018), ConLL-2003\textsuperscript{23} (Tjong Kim Sang and De Meulder, 2003), OntoNotes-5.0\textsuperscript{24} (Pradhan et al., 2013), EBMNLP\textsuperscript{25} (Nye et al., 2018) and BC2GM\textsuperscript{26} (Krallinger et al., 2015), SQuAD\textsuperscript{27} (Rajpurkar et al., 2016) and NewsQA\textsuperscript{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.

\textsuperscript{23}https://huggingface.co/datasets/conll2003
\textsuperscript{24}https://catalog.ldc.upenn.edu/LDC2013T19
\textsuperscript{25}https://huggingface.co/datasets/bigbio/ebm_pico
\textsuperscript{26}https://huggingface.co/datasets/bc2gm_corpus
\textsuperscript{27}https://rajpurkar.github.io/SQuAD-explorer
\textsuperscript{28}https://www.microsoft.com/en-us/research/project/newsqa-dataset/download/
list all the takeoffs and landings at general mitchell international

BackTrans

List of all takeoffs and landings at general mitchell international

EDA

list all the takeoffs and landings at general mitchell international

SSSBA

list all the compilers and events at general, international

AEDA

list all the takeoffs and landings at general? mitchell international!

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).

Iraq's President Saddam Hussein meets with chairman of the Russian liberal democratic party Vladimir Zhirinovsky.

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 Zhirinnovskly 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).

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.

BackTrans

Question: What solo album did Beyonce release in 2003?

EDA

Question: What single album was published by Beyonce in 2003?

SSSBA

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 Beyonc’s 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).
<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence</th>
<th>Instruction 1</th>
<th>Instruction 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yahoo</td>
<td>Shops in most malls advertise for Christmas help up to the last minute.</td>
<td>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 &amp; 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.</td>
<td>Write a brief document with a single sentence or multiple sentences corresponding to the following abstract description: &quot;Christmas help wanted ads in malls often run until the last minute.&quot;. 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 &amp; 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.</td>
</tr>
<tr>
<td>OTS</td>
<td>We are not obligated to publish any information or content on our Service and can remove it with or without notice.</td>
<td>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 PROP NOUN. 4. The document should have a length of 21-31 words. 5. Any sentence in the document should not include the abstract concept account management. 6. Any sentence in the document should not include the abstract concept jurisdiction.</td>
<td>Write a brief document with a single sentence or multiple sentences corresponding to the following abstract description: &quot;We reserve the right to remove content without notice.&quot;. Additionally, the document should have the following constraints: 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 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 account management. 6. Any sentence in the document should not include the abstract concept jurisdiction.</td>
</tr>
<tr>
<td>CoNLL-2003</td>
<td>Israel approves Arafat’s flight to West Bank.</td>
<td>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.</td>
<td></td>
</tr>
<tr>
<td>BC2GM</td>
<td>Comparison with alkaline phosphatases and 5 - nucleotidase</td>
<td>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.</td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Instruction prompts for various tasks.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sentence</th>
<th>Instruction 1</th>
<th>Instruction 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQUAD</td>
<td>Beyoncé's first solo recording was a feature on Jay Z's &quot;'03 Bonnie &amp; Clyde&quot; 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, &quot;Crazy in Love&quot;, featuring Jay Z, became Beyoncé’s first number-one single as a solo artist in the US. The single &quot;Baby Boy&quot; also reached number one, and singles, &quot;Me, Myself and I&quot; and &quot;Naughty Girl&quot;, both reached the top-five. The album earned Beyoncé a then record-tying five awards at the 46th Annual Grammy Awards; Best Contemporary R&amp;B Album, Best Female R&amp;B Vocal Performance for &quot;Dangerously in Love 2&quot;, Best R&amp;B Song and Best Rap/Sung Collaboration for &quot;Crazy in Love&quot;, and Best R&amp;B Performance by a Duo or Group with Vocals for &quot;The Closer I Get to You&quot; with Luther Vandross.</td>
<td>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, &quot;Her first solo album Dangerously in Love was released on June 24, 2003, after Michelle Williams and Kelly Rowland had released their solo efforts&quot;. 2. The document should have a length of 113-340 words.</td>
<td></td>
</tr>
</tbody>
</table>