CorrSynth - A Correlated Sampling Method for Diverse Dataset Generation from LLMs

Suhas S Kowshik*, **Abhishek Divekar***, **Vijit Malik**
Amazon
{kowssuhp, adivekar, vijitvm}@amazon.com

Abstract

Large language models (LLMs) have demonstrated remarkable performance in diverse tasks using zero-shot and few-shot prompting. Even though their capabilities of data synthesis have been studied well in recent years, the generated data suffers from a lack of diversity, less adherence to the prompt, and potential biases that creep into the data from the generator model. In this work, we tackle the challenge of generating datasets with high diversity, upon which a student model is trained for downstream tasks. Taking the route of decoding-time guidance-based approaches, we propose CorrSynth, which generates data that is more diverse and faithful to the input prompt using a correlated sampling strategy. Further, our method overcomes the complexity drawbacks of some other guidance-based techniques like classifier-based guidance. With extensive experiments, we show the effectiveness of our approach and substantiate our claims. In particular, we perform intrinsic evaluation to show the improvements in diversity. Our experiments show that CorrSynth improves both student metrics and intrinsic metrics upon competitive baselines across four datasets, showing the innate advantage of our method.

1 Introduction

Pretrained language models (LLMs) (Devlin et al., 2019) have achieved strong performance on text classification with a large amount of task-specific training data. However, in real world scenarios, collecting labeled data can be challenging due to expense and need for domain expertise. Recently, several works have focused on generating texts using versatile LLMs such as GPT-4 (Achiam et al., 2023), Claude (Bai et al., 2022), Mistral (Jiang et al., 2023), Mixtal (Jiang et al., 2024) and subsequently distill a student model on the synthetically generated data (West et al., 2022). However, generated datasets suffer from a lack of diversity (Yu et al., 2023a) and regurgitate the biases of the teacher LLMs, which proliferate into the student model. Although prior works have utilized retrieval augmented generation for diverse dataset synthesis (Divekar and Durrett, 2024), here we focus on the more fundamental challenge of improving or controlling generations *given a prompt and context*.

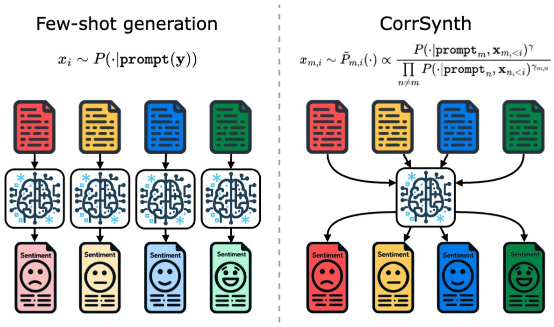


Figure 1: CorrSynth introduces anti-correlation between examples, compared to few-shot generation.

In particular, we focus on synthetic data generation for supervised text classification tasks and take the route of decoding time guidance based approaches (Sanchez et al., 2023; O'Brien and Lewis, 2023; Li et al., 2023; Chuang et al., 2023), which aim to tackle the challenge of improving diversity and faithfulness to target class in these generated datasets. Motivated by recent works on Classifier Free Guidance (CFG) (Sanchez et al., 2023), we introduce a novel guidance based strategy, CorrSynth. In CorrSynth, generations are kept faithful to the synthesis instruction, while introducing greater diversity and similarity to human text. CorrSynth is a correlated sampling approach which generates multiple sequences in parallel with strong inter-dependence between them. The main idea is as follows: when generating an instance of a particular class and sampling the next token, we *contrast* its logits with logits corresponding to partially generated instances from

*Equal contribution: order was determined by random dice rolls. Correspondence to: adivekar@amazon.com

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 16076–16095
November 12-16, 2024

other classes. This is a simple but crucial change compared to CFG: in CORRSYNTH, the contrasting logits for a class/label are obtained from generations corresponding to other labels, whereas in CFG, the contrasting logits are obtained feeding back the generation for the current label into the LLM with prompts corresponding to other labels. To synthesize a K–class classification dataset, this requires K–times fewer forward passes compared to CFG. Furthermore, we can smoothly trade-off diversity and improve class-separability by introducing contrasts between logits from the same or different classes.

In summary, our contributions are: (1) we develop a general correlated sampling approach, CORRSYNTH, that can generate multiple correlated sequences in parallel from an LLM, by explicitly introducing *contrasts* between parallel generations during the sampling of each token, (2) we apply this to classification dataset synthesis, with the goal of improving diversity of synthetic generations, (3) we demonstrate how our method overcomes the limitations of CFG and controllable synthesis in regards to diversity and label-separation, (4) we benchmark our approach on tasks ranging from humor detection, sentiment analysis and topic classification in regional news headlines. Our intrinsic analysis find that CORRSYNTH generates datasets with higher representation of tail entities, lexical diversity and similarity to human text, and distillation accuracy of student models, compared to four state of the art baselines.

2 Background

Notation: For $n \in \mathbb{N}$, let $[n] = \{1, 2, \cdots, n\}$. An LLM is defined through its vocabulary $\mathcal{V}$ and the auto-regressive sequence distribution P or equivalently the logits $\texttt{lg}$. Let $\mathcal{V}^* = \cup_{n\geq 1}\mathcal{V}^n$ denote the space of all finite sequences of tokens from $\mathcal{V}$. We denote sequences of tokens from $\mathcal{V}$ using lower boldface letters like $\mathbf{u}, \mathbf{v}$. For any sequence of tokens $\mathbf{w} = (w_1, \cdots, w_n) \in \mathcal{V}^n$ from $\mathcal{V}$, and any $j \in [n]$, let $\mathbf{w}_{<j} = (w_1, \cdots, w_{j-1})$ if $j > 1$, else, it is an empty sequence. Similarly $\mathbf{w}_{\leq j} = (w_1, \cdots, w_j)$. For any two sequences $\mathbf{u}, \mathbf{v} \in \mathcal{V}^*$ let $(\mathbf{u}, \mathbf{v})$ denote their concatenation. We denote by $P(\mathbf{v}|\mathbf{u})$ the conditional probability of generating $(\mathbf{u}, \mathbf{v})$ given that $\mathbf{u}$ has already been generated i.e., probability that $\mathbf{v}$ is a continuation of $\mathbf{u}$ for a given $\mathbf{u}$. Furthermore, for any $\mathbf{u}, \mathbf{v} \in \mathcal{V}^*$, we use $P(\cdot|\mathbf{u}, \mathbf{v})$ to denote the conditioning on the concatenation $(\mathbf{u}, \mathbf{v})$. For any prompt $\texttt{prompt} \in \mathcal{V}^*$, and any $\mathbf{w} \in \mathcal{V}^n$, the auto-regressive distribution P satisfies

$$P(\mathbf{w}|\texttt{prompt}) =$$

$$P(w_1|\texttt{prompt}) \prod_{j=2}^{n} P(w_j|\texttt{prompt}, w_1, \cdots, w_{j-1})$$

When we describe natural language domains using $\mathcal{X}, \mathcal{Y}$ we mean either in the sense of them containing natural language sentences or as subsets of $\mathcal{V}^*$, it will be clear from the context.

We consider dataset generation for text classification tasks. Suppose we have a multiclass text classification problem with K classes as $[K]$ and input domain $\mathcal{X}$. Let $\mathcal{Y} = \{\mathbf{y}_1, \cdots, \mathbf{y}_K\}$ be the space of label verbalizations for the K classes i.e., $\mathbf{y}_k$ is a textual description of label $k \in [K]$. A natural language example input is denoted as $\mathbf{x} \in \mathcal{X}$. So the learning problem is defined on $\mathcal{X} \times \mathcal{Y}$: given a data generating distribution P_{XY} on $\mathcal{X} \times \mathcal{Y}$ the task is to learn a classifier $h : \mathcal{X} \to \mathcal{Y}$ (using some training data) such that $\mathbb{E}\left[l(h(\mathbf{x}), \mathbf{y})\right]$ is minimized for a given loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where the expectation is taken with respect to P_{XY}.

Given the rapid advancement of LLMs like GPT-4, Llama2, Mistral etc. we are interested in utilizing the world knowledge and reasoning capabilities of these large models to generate synthetic training data for the textual K-class classification problem. Similar to recent works in this domain (Ye et al., 2022a; Gao et al., 2022; Meng et al., 2022a, 2023a; Yu et al., 2023b; Ye et al., 2022c; Yu et al., 2024; Guo and Chen, 2024), we consider the setup of prompting teacher LLM with a prompt $\texttt{prompt}$ that includes a label $\mathbf{y} \in \mathcal{Y}$, a few In-Context Learning (ICL) examples for the label $\mathbf{y}$ and potentially any other instance dependent attributes, and the prompt tasks the LLM to generate a synthetic instance $\mathbf{x} \in \mathcal{X}$ whose true label is expected to be $\mathbf{y}$ i.e., the aim is to generate $x \sim P_{X|Y=\mathbf{y}}$. That is, we generate a synthetic dataset $\mathcal{D}_{\text{SYNTH}}$. A student language model (e.g., a BERT-style pre-trained encoder model (Devlin et al., 2019)) is trained on $\mathcal{D}_{\text{SYNTH}}$.

For the ICL examples, we assume that we have access to a *seed set* of examples $\mathcal{D}_{\text{SEED}} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$. For us, typically n is such that we have around 50 examples per class. We assume that $\mathcal{D}_{\text{SEED}}$ is not large enough to train an effective student, but instead a larger synthetic dataset $\mathcal{D}_{\text{SYNTH}} = \{(\tilde{\mathbf{x}}_i, \mathbf{y}_i)\}_{i=1}^{m}$ will be needed.

A standard approach to dataset synthesis is few shot generation i.e. FewGen (Brown et al., 2020a; Ye et al., 2022c; Yehudai et al., 2024). For instance, consider a task of detecting a business news article. In order to synthesize a dataset for this task, we could prompt the LLM appropriately, include few ICL examples. The LLM might generate a fairly decent article. But when we sample a large number of generations we see that the there is lack of diversity: similar entities are repeated, popular topics are highlighted and potential stylistic differences from a human written text. These could affect the performance of a student model that is trained on such dataset.

A "good" synthetic dataset must ensure that the conditional distribution of instances given any label must closely approximate that of the true distribution P_{XY}. This includes: i) correct semantic separation of labels, ii) preservation of intra-label semantic diversity and of course, iii) fluent and coherent generations. In order to achieve (i) and (ii) (without compromising on (iii)), we present a method, CorrSynth, in the flavor of decoding time guidance techniques (Li et al., 2023; O'Brien and Lewis, 2023; Sanchez et al., 2023; Chuang et al., 2023). In these works, at inference time, the token probability distribution is tilted by another distribution obtained either from a different LLM, or same LLM with a different prompt, or different layers of the same LLM. In particular, we take inspiration from the classifier free guidance (Ho and Salimans, 2021) method applied to text based LLMs (Sanchez et al., 2023). CorrSynth aims to control i) diversity in generations, ii) similarity to human crafted gold dataset, iii) cross label separation and at the same time iv) improve the student performance. The core idea of our approach is to perform correlated or dependent sampling from the LLM i.e., multiple sequences are generated in parallel that have strong dependency between each other. Figure 1 illustrates our method. More details are given in section 3. This method can be used in conjunction with other synthetic dataset generation approaches like retrieval augmented generation (Lewis et al., 2020).

3 Method

Now we describe our novel CorrSynth method of sampling from an LLM. Although it is a general technique, we choose to motivate it from the perspective of data synthesis for a text based supervised learning problem.

3.1 CorrSynth

Let us consider the case of binary classification with verbalized labels $\{\mathbf{y}_0, \mathbf{y}_1\}$. As is standard in dataset synthesis (Ye et al., 2022a; Brown et al., 2020b), we create class-conditioned prompt $\texttt{prompt}(\mathbf{y})$ which describes the task using verbalization $\mathbf{y} \in \{\mathbf{y}_0, \mathbf{y}_1\}$, and prompt the LLM to generate continuations as our synthetic input x. In-context examples are used to guide the generations to follow the format specified in the prompt. Suppose we want to generate two instances $\mathbf{x}, \bar{\mathbf{x}}$ corresponding to labels $\mathbf{y}, \bar{\mathbf{y}}$ respectively. In CorrSynthwe generate them together as follows. Let $0 \leq \delta \leq \gamma$. Then:

$$x_i \sim \tilde{P}_i(\cdot) \propto \frac{P(\cdot|\texttt{prompt}(\mathbf{y}), \mathbf{x}_{<i})^{\gamma}}{P(\cdot|\texttt{prompt}(\bar{\mathbf{y}}), \bar{\mathbf{x}}_{<i})^{\gamma-\delta}} \quad (1)$$

$$\bar{x}_i \sim \tilde{Q}_i(\cdot) \propto \frac{P(\cdot|\texttt{prompt}(\bar{\mathbf{y}}), \bar{\mathbf{x}}_{<i})^{\gamma}}{P(\cdot|\texttt{prompt}(\mathbf{y}), \mathbf{x}_{<i})^{\gamma-\delta}} \quad (2)$$

We hypothesize that the sequences $\mathbf{x}, \bar{\mathbf{x}}$ generated auto-regressively using equations (1) and (2) are naturally anti-correlated: they tend to be far apart in the embedding space of the LLM. This is because, when sampling a token for a sequence, the plausible tokens for the contrasting sequences are weighted down. Furthermore, at token i, even if the numerator and denominator distributions in (1) highlight different entities or parts of speech, we expect the overall semantic meaning to be weakly present in the individual token distributions due to the attention mechanism. Thus even at these tokens, we posit that the contrast provides a signal that moves the generated sequences apart. This reasoning is based on intuition that requires careful experiments to prove. Nonetheless, we will demonstrate this separation of sequences in our analysis in section 6. So we call the sequences $\mathbf{x}, \bar{\mathbf{x}}$ to be contrastive to one another. We can use this property to control label separation as well as intra-class diversity when generating synthetic instances.

Crucial change from CFG: in denominator of (1), the conditioned partial sequence $\bar{\mathbf{x}}_{<i}$ is actually expected to be faithful to $\texttt{prompt}(\bar{\mathbf{y}})$, and thus the effect of guidance would persist even after many tokens. Additionally, we generate two sequences together, leading to a two fold increase in the number of forward passes compared to a single generation, whereas CFG would require four times more. We introduce another parameter δ which controls the

strength of the denominator contrast. More details on CFG for dataset synthesis are in Appendix D.

3.2 M–CorrSynth

Next, we generalize from binary to M-way contrastive generation. Suppose we have M prompts $\{\texttt{prompt}_1, \cdots, \texttt{prompt}_M\}$. We want to generate M sequences $\{\mathbf{x}_m : m \in [M]\}$ such that $\mathbf{x}_m$ is faithful to $\texttt{prompt}_m$. Let $\gamma > 0$ be the guidance, and let $0 \leq \delta \leq \gamma$. We introduce M^2 weights $\{\gamma_{m,n} : m, n \in [M], \gamma_{m,m} = 0\}$. We generate the i-th token of $\mathbf{x}_m = (x_{m,1}, \cdots, x_{m,n_m}), \forall m$:

$$x_{m,i} \sim \tilde{P}_{m,i}(\cdot) \propto \frac{P(\cdot|\texttt{prompt}_m, \mathbf{x}_{m,<i})^{\gamma}}{\prod_{n \neq m} P(\cdot|\texttt{prompt}_n, \mathbf{x}_{n,<i})^{\gamma_{m,n}}} \quad (3)$$

Next we describe our choice of $\gamma_{m,n}$.

3.2.1 Uniform contrastive guidance

We set a parameter δ that controls the total amount of contrast guidance: for each m, $\sum_n \gamma_{m,n} = \gamma - \delta$. Then, when generating i-th token for $\mathbf{x}_m$, we set $\gamma_{m,n} = 0$ for sequences $\mathbf{x}_n$ that have already hit the EOS token. Then, we uniformly divide $\gamma - \delta$ among remaining $\gamma_{m,n}$[1]. More details are in Appendix E.1. Using uniform contrastive guidance, M-CorrSynth has a natural geometric mean interpretation that we discuss in Appendix E.

3.2.2 CorrSynth for K-class synthesis

Now we briefly describe how we use CorrSynth in data synthesis for K class classification. Recall that in K-class classification problem over $\mathcal{X} \times \mathcal{Y}$ we have classes $[K]$ with label verbalizations $\{\mathbf{y}_1, \cdots, \mathbf{y}_K\}$. To generates instances for each class, we create prompts as follows. Let $R \in \mathbb{N}$ be the repeat factor. In M-CorrSynth, we take $M = KR$, and prompts in $\{\texttt{prompt}_m : m = (k-1)R + r, 1 \leq r \leq R\}$ correspond to class k for all $k \in [K]$. For $m = (k-1)R + r$, prompt $\texttt{prompt}_m$ asks the LLM to generate instances for class k contains positive ICL examples for that class. These ICL examples differ across r. Thus in equation (3), a generation for class k is, potentially, contrasted against the remaining $R - 1$ generations from the same class, as well as the $(K-1)R$ generations from other classes. Based on setting the weights $\gamma_{m,n}$ to be zero for either intra-label terms or cross label terms, we get three scenarios:

CorrSynth Cross-label: When generating a sequence for class k and $m = (k-1)R + r$, we set $\gamma_{m,n} = 0$ for $n \in \{(k-1)R + r' : r' \neq r\}$. So only terms belonging to classes $k' \neq k$ appear in the denominator of (3).

CorrSynth Intra-label: When generating a sequence for class k and $m = (k-1)R + r$, we set $\gamma_{m,n} = 0$ for $n \in \{(k'-1)R + r' : r' \in [R], k' \neq k\}$. So only terms belonging to class k appear in the denominator of (3).

CorrSynth Hybrid: denominator of (3) contains terms that belong to the same class as well as those that belong to other classes. We separately set the target guidance for each of the Cross- and Intra-label terms: we fix two targets γ_{intra} and γ_{cross} such the sum of $\gamma_{m,n}$ for Intra and Cross label terms are set to γ_{intra} and γ_{cross} respectively. Then we uniformly split the target guidances γ_{intra} and γ_{cross} in respective groups. More details of K-class CorrSynth is given in Appendix E.3

Dataset	Type	Class	Train, Test
AG News	Topic	4	115K, 7.6K
ToI Headlines	Topic	10	52K, 10K
Humor	Sentiment	2	15K, 3K
IMDb	Sentiment	2	20K, 25K

Table 1: Dataset statistics.

3.2.3 Logits Space computation

The CorrSynth method is implemented using vector arithmetic in the space of LLM outputs i.e. logits space. Complete details are in Appendix E.4. Taking logarithm of the CorrSynth sampling equations gives us similar results[2].

3.2.4 Plausibility Constraint (α)

The contrast terms in CorrSynth could sometimes upweight irrelevant tokens i.e. those which are not plausible conditioned on the prompt/label under consideration. To mitigate this, we borrow the idea of plausibility constraint from (Li et al., 2023; O'Brien and Lewis, 2023) to limit the token up weighting space: by reducing the space of logits to those tokens that have at least α fraction of the mode of the numerator distribution in (3). We provide the complete formulation in Appendix E.5.

4 Experimental Setup

Datasets. We experiment on 4 datasets described in Table 1, which are selected to encompass a wide scope of generation tasks (news headlines, news

[1] $\gamma_{m,n}$ also depends on the token index i; we suppress it.

[2] Caveat: taking logarithm gives us log-probabilities which are normalized version of logits. Experimentally, we have not found significant impact of this normalization.

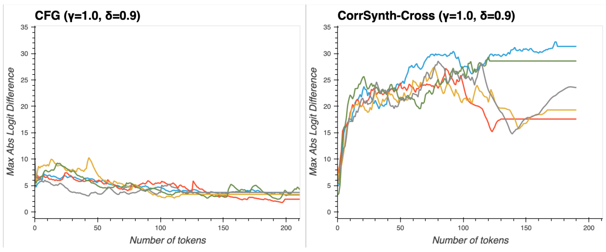


Figure 2: Generation progression from CFG and CorrSynth. We sample five generations using 3-shot prompts from IMDb. The colored lines represent the absolute difference between logits of the current generation and contrast for each generation timestep (taken as an exponential moving average).

articles, humorous product questions and movie reviews). Previous work primarily benchmarked only sentiment and topic classification datasets. We consider: (1) AG News (Zhang et al., 2015), a popular topic classification dataset where each news summary is mapped to a news topic. The generation task involves generating news summaries based on news topics; (2) ToI Headlines(Kulkarni, 2020), similarly is a topic classification dataset of regional news headlines in India that maps news topics to news headlines; the generation task is similar to AG News. The difficulty is that the headlines is regionalized to Indian news and hence requires India specific entities; (3) Humor (Ziser et al., 2020) task involves generating humorous and non-humorous questions from retrieved product details; (4) IMDb (Maas et al., 2011) is a sentiment task with binary labels. Prompts are in Appendix G.

Teachers and students. As a teacher model, we use a frozen Mixtral (8x7B) (Jiang et al., 2024) or Phi-3 Mini (3.8B) (Abdin et al., 2024) for the data generation step. Following (Divekar and Durrett, 2024), we select examples randomly from the train set: 50 ICL examples per class for multi-class and 100 per class for binary. We think that this is a reasonable number of labeled examples since we are trading off the effort of labeling versus developing a potential zeroshot technique (which may not work well in practice). We use DistilBERT student model (66M params Sanh et al. (2019)) as it is popular in prior work.

Evaluation criteria The task of evaluation of quality of text generation is quite challenging (Chang et al., 2024). Following prior works like (Divekar and Durrett, 2024), we evaluate synthetic generations based on several metrics. **Self-BLEU** (Papineni et al., 2002; Zhu et al., 2018) measures lexical diversity of a corpus of texts based on n-gram overlap between pairs of examples. **Entity entropy** measures the *diversity* of entities in the generated texts using the distribution of each of 16 entity-types (inferred from a pre-trained named entity recognition model). Dataset which have high occurrence of few popular entities score lower on entropy. **MAUVE** (Liu et al., 2021) measures closeness to human-written text using representations from a pre-trained GPT2-XL model. We also measure the **student accuracy** when trained on the synthetic data. We do not consider label preservation accuracy as it is susceptible to easy examples (Divekar and Durrett, 2024). In order to analyse the behavior of our strategy, we also study the label-wise cosine similarity of the generations, low dimensional embeddings of the generations using UMAP (McInnes et al., 2020) and dataset cartography (Swayamdipta et al., 2020).

Remark on diversity In this work we are concerned about diversity at a dataset level and not an an instance level. To illustrate the difference between these two, consider the task of generating a long story. Here, it is important to ensure that the generated story has many of the features of a human written story (like complex storyline, many characters, non-repeating scenes etc.). But notice that ensuring such an instance level diversity does not guarantee diverse dataset of stories: multiple such stories obtained from an LLM could have a lot of overlap in content. For synthesis of good classification datasets, we require a more global notion of diversity which is at the dataset level.

5 Results

5.1 Comparison to CFG

We compare the effect of contrast as next-token generation proceeds in CFG and CorrSynth. To this end, we consider IMDb, and sample continuations for five 3-shot prompts from both CFG and CorrSynth for the same Cross-label parameters: $\{R = 1, \gamma = 1.0, \delta = 0.9, \alpha = 0\}$. For each token, we store the maximum absolute difference of the current label logits vector and the contrast label logits vector (i.e. ∞-norm of logits difference of numerator and denominator in (1) for CorrSynth, and similar terms in CFG). We plot this difference against the tokens generated.

Figure 2 shows the difference between CFG and CorrSynth: as token generation proceeds, the effect of contrast in CFG is muted. This happens

Method	Teacher LM	Accuracy (↑) AG.	ToI	Hum.	IMDb	Avg.	MAUVE (↑) AG.	ToI	Hum.	IMDb	Avg.
Gold	-	91.4	78.9	92.9	91.4	88.7	-	-	-	-	-
In-Context Learning											
FewGen	Phi-3 Mini	83.8	69.7	68.5	85.1	76.8	91.0	86.3	83.7	67.7	82.2
FewGen	Mixtral	72.3	47.3	82.8	87.1	67.5	87.1	91.6	87.0	64.6	82.6
Corr-Intra	Phi-3 Mini	84.8	71.0	84.7	87.1	81.9	82.3	83.2	82.3	77.4	81.3
Corr-Hybrid	Phi-3 Mini	**85.1**	**71.1**	85.1	86.8	**82.1**	77.5	82.0	81.7	71.0	78.1
Corr-Intra	Mixtral	78.5	68.9	**86.5**	**88.6**	80.1	**94.4**	95.6	95.5	76.8	90.1
Corr-Hybrid	Mixtral	73.6	68.4	86.0	88.1	79.0	93.8	**96.1**	**97.1**	**80.5**	**91.9**

Method	Teacher LM	Self-BLEU-5 (↓) AG.	ToI	Hum.	IMDb	Avg.	Entity-Entropy (↑) AG.	ToI	Hum.	IMDb	Avg.
Gold	-	17.1	7.9	19.8	27.9	18.2	6.6	6.1	5.1	7.5	6.3
In-Context Learning											
FewGen	Phi-3 Mini	33.9	15.3	39.9	57.7	36.7	6.6	6.3	4.3	5.3	5.6
FewGen	Mixtral	39.4	37.9	64.6	66.5	52.1	5.9	5.2	3.6	5.2	5.0
Corr-Intra	Phi-3 Mini	13.1	9.0	23.5	24.9	17.6	**7.4**	**6.9**	**4.9**	**6.5**	**6.4**
Corr-Hybrid	Phi-3 Mini	**12.1**	**8.7**	**22.8**	**19.2**	**15.7**	**7.4**	**6.9**	4.8	6.4	**6.4**
Corr-Intra	Mixtral	18.9	17.6	45.3	33.0	28.7	6.3	5.7	3.7	6.0	5.4
Corr-Hybrid	Mixtral	17.5	18.4	41.4	27.4	26.2	6.5	5.6	4.1	6.4	5.7

Table 2: Evaluation of intrinsic dataset quality and DistilBERT student model fine-tuned on real and synthetic datasets. We report mean accuracy numbers across 5 runs. When generating each instance, we select 3 in-context examples at random to prime the LLM's next-token distribution before sampling continuations.

since the same generated sequence for the current label is fed back into the contrast model and thus the effect of the contrastive prompt reduces over later token generations. Whereas in CorrSynth, the effect of the guidance or contrast persists. As a result, we believe CorrSynth is a better suited for longer generations where guidance is required for the entirety of generation. In terms of complexity, as discussed previously, we incur a much higher complexity of LLM model forward passes in CFG (detailed comparison in Appendix F.1).

5.2 Comparison to FewGen

In this section, we present our experimental results against FewGen. We use the following settings:

CorrSynth Cross-label: Repeat factor $R = 1$, Uniform contrastive guidance with $\gamma = 1.0$ and $\delta = 0.9 \times \gamma$ and plausibility criterion $\alpha = 10^{-3}$.

CorrSynth Intra-label: Repeat factor $R = 2$, Uniform contrastive guidance with $\gamma = 1.0$ and $\delta = 0.5 \times \gamma$ and plausibility criterion $\alpha = 10^{-3}$.

CorrSynth Hybrid: Repeat factor $R = 2$, set $\gamma = 1.0$, Set $\gamma_{intra} = \gamma/2$, $\gamma_{cross} = \gamma/10$. Then uniform contrastive guidance in each of intra and cross terms. We set plausibility criterion $\alpha = 10^{-3}$.

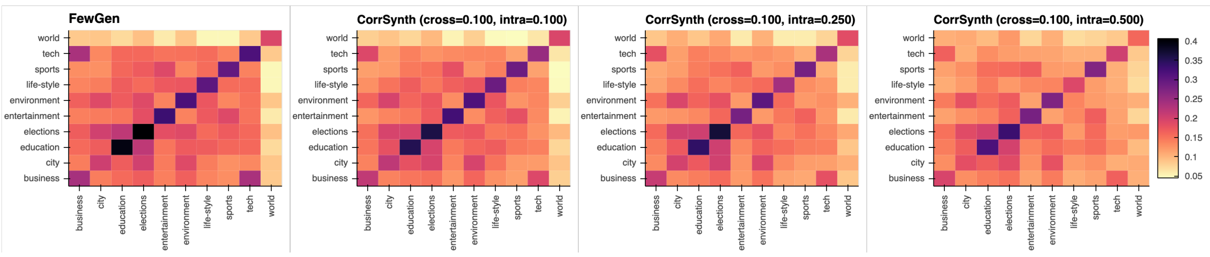
We observe in Table 2 that 3-shot CorrSynth outperforms FewGen on all evaluation metrics. Specifically, using Hybrid and Intra variants, we can achieve better student model accuracy (DistilBERT) while increasing diversity (lower Self-BLEU, higher entity entropy) and better match with human-written text (better MAUVE). For MAUVE computation, we have used embeddings based on a GPT-2XL model. We have only shown the results for Intra and Hybrid variants since from our ablations they performed best. In Appendix B, we note the zero-shot results, which demonstrate comparable gains on all metrics.

5.3 Comparison to prior works

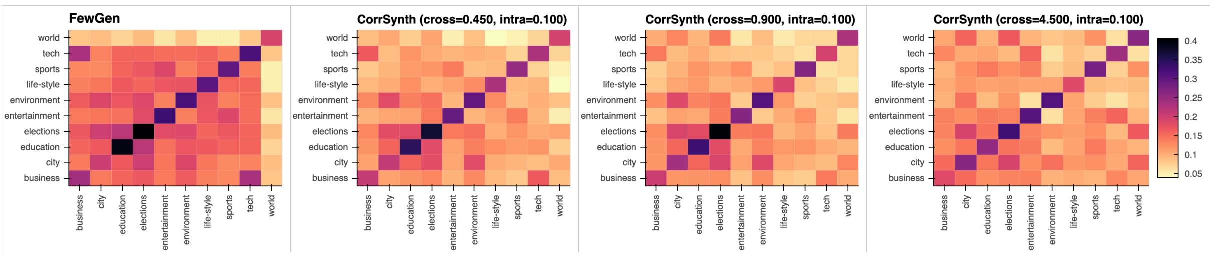
In Table 3 we compare CorrSynth to current dataset generation methods as baselines. Base-

Method	Teacher LM	Accuracy		MAUVE		Self-BLEU-5		Entity-Entropy	
		AG.	IMDB	AG.	IMDB	AG.	IMDB	AG.	IMDB
GOLD	-	91.4	91.4	-	-	17.1	27.9	6.6	7.5
RETRIEVAL-BASED METHODS									
REGEN	BERT	82.7	⊗	68.1	⊗	56.5	⊗	**8.1**	⊗
SYNTHESIZRR	LLAMA2	84.6	84.8	**92.6**	72.6	34.2	62.9	7.2	5.7
NON-RETRIEVAL METHODS									
SUNGEN	GPT2-XL	⊗	84.9	⊗	68.7	⊗	**15.4**	⊗	4.9
S3	GPT3.5-T	⊗	**87.1**	⊗	62.0	⊗	62.2	⊗	5.7
ATTRPROMPT	GPT3.5-T	79.8	⊗	52.8	⊗	39.8	⊗	6.0	⊗
(Ours) CORR-Intra	PHI-3 MINI	84.8	**87.1**	82.3	**77.4**	13.1	24.9	7.4	**6.5**
(Ours) CORR-Hybrid	PHI-3 MINI	**85.1**	86.8	77.5	71.0	**12.1**	19.2	7.4	6.4

Table 3: Comparison of quality metrics and DISTILBERT student model fine-tuned on 6k rows from each approach. Mean accuracy across 5 training runs is considered. ⊗ indicates datasets were not released by authors.



(a) Impact of increasing Intra-label contrast (left to right) in Hybrid CORRSYNTH upon label-wise cosine similarities.



(b) Impact of increasing Cross-label contrast (left to right) in Hybrid CORRSYNTH upon label-wise cosine similarities.

Figure 3: Heatmaps for label-wise cosine similarities on TOI HEADLINES (with Phi-3-mini) as we increase Intra-label contrast vs increasing cross-label contrast. Note that "Cross" and "Intra" in figure titles correspond to γ_{cross} and γ_{intra} respectively. FEWGEN heatmaps are provided for reference.

line numbers are quoted from Divekar and Durrett (2024), where all results are reported on 6k rows using DISTILBERT student (same as our setup). The following SOTA generation methods have been compared: (1) **REGEN** (Yu et al., 2023c): uses dual BERT models - one for retrieval and one as a classifier - to perform multi-round filtering and eliminate noisy data based on model consistency; (2) **SYNTHESIZRR** (Divekar and Durrett, 2024): develops a hybrid retrieval-augmentation based approach to rewrite contexts, greatly enhancing the diversity of generated text; (3) **SUNGEN** (Gao et al., 2023): employs ZEROGEN (Ye et al., 2022a) to create a substantial synthetic dataset (200k rows) and then uses a bi-level optimization algorithm to assign instance-weights to each synthetic example; (4) **S3** (Wang et al., 2023a): builds a distinct "seed dataset" to train a student model, leverages an LLM to identify errors, and synthesizes supplementary data. This cycle of data augmentation is repeated. (5) **ATTRPROMPT** (Yu et al., 2023a): enhances dataset diversity and unbiasedness by prompting a potent LLM like GPT3.5-TURBO with attributes identified through human-in-the-loop task analysis.

We divide our comparison into non-retrieval and retrieval based synthesis, as the latter naturally demonstrates higher diversity (Divekar and Durrett, 2024). We observe that CORRSYNTH achieves strong performance on all metrics, despite using a small teacher LLM (PHI-3 MINIwith 3.8B parameters) compared to prior approaches.

6 Analysis and Visualizations

Effect of Intra-label and Cross-label contrasts: Given the promising results of our method CORRSYNTH, we wanted to analyse and visualize the effect of varying Intra-label and cross-label contrasts upon the generations. For this, we obtain the average label-wise cosine similarities of the generations and plot them as heatmaps (see Figure 3). We specifically study the behavior of our approach in multi-label setting upon TOI HEADLINES dataset to emphasize our results. In practice, we use the `all-mpnet-base-v2` model from SentenceTransformers library to obtain the text representations of each generation. Next, for generations each label i and j in the labelspace, we compute the pairwise cosine similarities of all generations corresponding to label i, to that of label j. The pairwise cosine similarities are then added and the mean label-wise similarity is computed. We hypothesize that in our approach (Hybrid CORRSYNTH), if the Intra label contrast is increased, then, within each class the generations should get further away so that the net cosine similarity within the class should reduce across all classes. As we can see in Figure 3a, the diagonals become lighter as we go left to right. On the other hand, if the cross-label contrast is increased net separability between every pair of classes should increase. As expected, we can see in Figure 3b, in the heatmaps the off-diagonal label similarities are becoming lighter as cross contrast is increased.

Effect of δ: To visualize the effect of varying δ on the generations obtained through CORRSYNTH, we plot 2-dimensional representations of the generations. We use Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) for Dimension Reduction[3] of text representations obtained using `all-mpnet-base-v2` model. As earlier, we perform this analysis in a multi-label setting with TOI HEADLINES.

In Figure 4, we can see that as δ is reduced from $(0.9, 0.5, 0)$, the representations become more and more diffused with each other, leading to overlaps, making the student model hard to learn the decision boundary. For $\delta = 0.9$, we can visualize the clusters containing data points from different labels are well-separated, which resonates with our best performing results as well. Note that overlapping/diffused datapoints could be indicators of mislabelled generations as well as hard negatives.

We hypothesize that as we decrease delta from 0.9, first we see an increase in hard negative generations than mislabeled generations, whereas after some threshold, the extent of mislabeled generations increase. Thus there is a sweet spot which provides good amount of hard examples with minimal number of wrong generations. We can see this effect in the corresponding cartography plots (Swayamdipta et al., 2020) in Figure 5 where as we go from left to right, the density of gray and blue points increase but blue points density increases more for much smaller delta than for gray points. The gray points here typically denote hard to learn examples, where as the blue one predominantly represent mislabeled example. These hard negative generations benefit the student model training.

7 Related Work

Dataset synthesis using LLMs. In recent years LLMs have exhibited strong generative capabilities (Brown et al., 2020a; Cobbe et al., 2021) to solve a diverse range of tasks. With well-designed prompts, large-scale LLMs have shown its notable zero-shot and few-shot learning ability (Shin et al., 2020; Jiang et al., 2020; Reynolds and McDonell, 2021).

More recently, these models have gained popularity in their superior ability to synthesize task-specific datasets (Wang et al., 2021; Lee et al., 2021; Kumar et al., 2020; Puri et al., 2020; Anaby-Tavor et al., 2019). LLMs such as GPT-3 (Wang et al., 2023b; Honovich et al., 2023; West et al., 2022) and chat-tuned models (Yehudai et al., 2024; Yu et al., 2023a; Wang et al., 2023a) have shown promising results on the task of generating synthetic data. Certain works like Meng et al. (2023b) fine-tune an LLM to generate NLU datasets, whereas our work is similar to Schick and Schütze (2021); Meng et al. (2022a) which use frozen LLMs with task-dependent prompts to generate data, targeting text classification.

Text classification dataset synthesis employs class-specific prompts; previous studies explored zero-shot (Ye et al., 2022b) and iterative few-shot prompting (Ye et al., 2022d). However, only recently has the lack of diversity in synthetic classifi-

[3] https://umap-learn.readthedocs.io/en/latest/

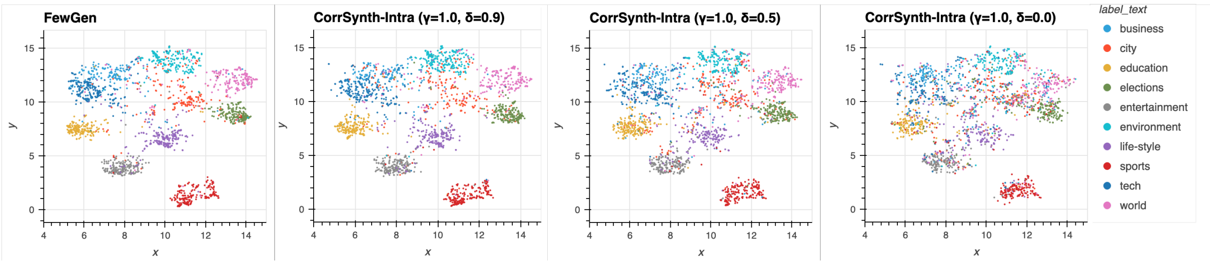


Figure 4: Visualising two-dimensional text representations of generations (on TOI HEADLINES with PHI-3 MINI) using CORRSYNTH-Intra. We gradually increase guidance delta, δ in $(0.0, 0.5, 0.9)$. FEWGEN plot is provided as a reference to the unmodified clusters (it is equivalent to $\delta = 1$ i.e. no contrast).

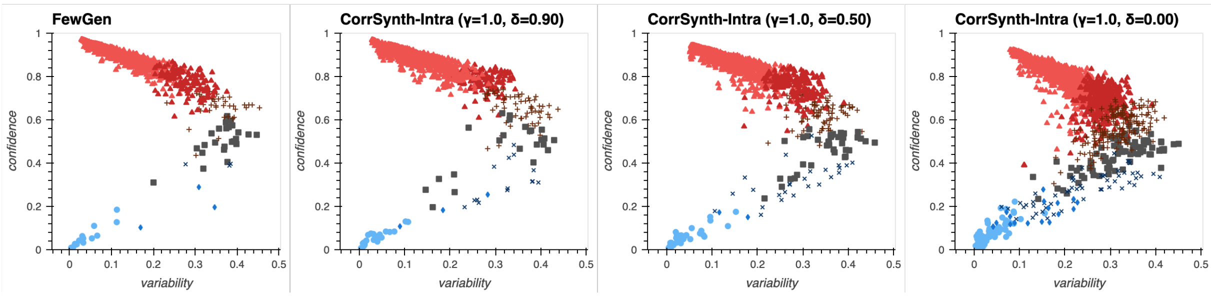


Figure 5: Datamaps for DISTILBERT training run on 2K examples of TOI HEADLINES generated using CORRSYNTH-Intra using Phi-3-mini. FEWGEN datamap is provided for reference.

cation datasets been recognized. Yu et al. (2023a) advocated for using diverse prompts that explicitly introduce variations, such as subtopics and brands, resulting in more diverse conditioning. In contrast, our approach achieves diversity with a fixed prompt. Divekar and Durrett (2024) employs retrieval augmentation to introduce variety into the dataset synthesis process by seeding the LLM with different content. However, the diversity here is constrained by the corpus availability, whereas our work improves diversity despite relying only on LLM parameters.

Classifier-Free Guidance (CFG) is a sampling technique introduced in diffusion literature (Ho and Salimans, 2021) and later extended to autoregressive LLMs (Sanchez et al., 2023). CFG falls under general guidance based techniques, where a guidance distribution is used at inference to alter the sampling distribution towards the desired goal.In CFG, this guidance distribution is provided by the LLM itself but with a different prompt as described in Appendix D. Context-aware decoding Shi et al. (2023) also uses the same formulation as CFG.

Contrastive decoding (CD refers to another family of guidance based methods that derive the guidance distribution from either a smaller LLM (O'Brien and Lewis, 2023; Li et al., 2023), different layers of the same LLM (Chuang et al., 2023; Gera et al., 2023). In all these methods from CFG to CD, the idea is essentially that to generate a sequence, a contrasting distribution is computed at inference. But different sequences are generated independently. In CORRSYNTH, although we borrow the general idea of a using a guidance-providing distribution at inference, the guidance distribution itself corresponds to a actual parallel generation providing both a) (anti-)correlation between multiple sequences as desired, b) compute efficiency. See section 3 and Appendix F.

8 Conclusion

In this work, we propose a novel technique CORRSYNTH which uses correlated sampling and intuition from classifier free guidance and contrastive decoding to generate strongly diverse datasets across a variety of tasks with good cross-label separations. We provide the mathematical intuition of our approach and back our theoretical discussion with empirical results. Our extensive experimentation across 4 datasets show the robustness of our approach in generating synthetic data.

In the future, we would like to study the effect of including Intra-label contrasts while generating with the LLMs, and mixing up both cross-label and Intra-label contrasts (a hybrid approach) to see how the generations are affected with respect to both intrinsic and extrinsic evaluation.

9 Limitations

The scope of our experiments is restricted to a set of classification tasks over a few English domains of text. While we believe our approach can be applied to other languages, other domains, and tasks like question answering that go beyond classification, we have not validated this in this work. Furthermore, the scope of our formulation is restricted to supervised learning problems where there a well-defined or natural label space. Extensions to unsupervised tasks like datasets for pre-training is an interesting possibility to be explored. The introduction of new hyper-parameters in any method requires tuning, which increases costs. In our case a high value of δ with respect to the original guidance γ (e.g. $\delta = 0.9 * \gamma$, yields positive results for all guidance values). However, the tuning of the initial guidance parameter was subject to a heuristic search. Finally, our approach performs modifications to the generation process by performing correlated sampling in the logits space. This makes our approach infeasible to use with API-only teacher LMs such as GPT-4, Claude, Gemini, etc.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, N. Tepper, and Naama Zwerdling. 2019. Do not have enough data? deep learning to the rescue! In *AAAI Conference on Artificial Intelligence*.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhishek Divekar and Greg Durrett. 2024. Synthesizrr: Generating diverse datasets with retrieval augmentation. *arXiv preprint arXiv:2405.10040*.

Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, WEIZHONG ZHANG, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2022. Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning. In *The Eleventh International Conference on Learning Representations*.

Jiahui Gao, Renjie Pi, LIN Yong, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-guided noise-free data generation for efficient zero-shot learning. In *The Eleventh International Conference on Learning Representations*.

Ariel Gera, Roni Friedman, Ofir Arviv, Chulaka Gunasekara, Benjamin Sznajder, Noam Slonim, and Eyal Shnarch. 2023. The benefits of bad advice: Autocontrastive decoding across model layers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10406–10420, Toronto, Canada. Association for Computational Linguistics.

Xu Guo and Yiqiang Chen. 2024. Generative AI for Synthetic Data Generation: Methods, Challenges and the Future. *arXiv preprint arXiv:2403.04190*.

Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Rohit Kulkarni. 2020. Times of India News Headlines.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Kenton Lee, Kelvin Guu, Luheng He, Timothy Dozat, and Hyung Won Chung. 2021. Neural Data Augmentation via Example Extrapolation. *ArXiv*, abs/2102.01335.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive Decoding: Open-ended Text Generation as Optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312.

Lang Liu, Krishna Pillutla, Sean Welleck, Sewoong Oh, Yejin Choi, and Zaid Harchaoui. 2021. Divergence Frontiers for Generative Models: Sample Complexity, Quantization Effects, and Frontier Integrals. In *Advances in Neural Information Processing Systems*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. pages 142–150, Portland, Oregon, USA.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. *Preprint*, arXiv:1802.03426.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022a. Generating training data with language models: Towards zero-shot language understanding. *ArXiv*, abs/2202.04538.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022b. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 462–477. Curran Associates, Inc.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023a. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR.

Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. 2023b. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. 2023. Stay on topic with classifier-free guidance. *arXiv preprint arXiv:2306.17806*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019*.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *ArXiv*, abs/2305.14739.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the*

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Ruida Wang, Wangchunshu Zhou, and Mrinmaya Sachan. 2023a. Let's synthesize step by step: Iterative dataset synthesis with large language models by extrapolating errors from small models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11817–11831, Singapore. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *ArXiv*, abs/2109.09193.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022b. Zerogen: Efficient zero-shot learning via dataset generation. *ArXiv*, abs/2202.07922.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022c. ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022d. ProGen: Progressive zero-shot dataset generation via in-context feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. *ArXiv*, abs/2401.14367.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023a. 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*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2024. Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems*, 36.

Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. 2023b. Regen: Zero-shot text classification via training data generation with progressive dense retrieval. *arXiv preprint arXiv:2305.10703*.

Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. 2023c. ReGen: Zero-shot text classification via training data generation with progressive dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11782–11805, Toronto, Canada. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. *SIGIR*.

Yftah Ziser, Elad Kravi, and David Carmel. 2020. Humor detection in product question answering systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 519–528, New York, NY, USA. Association for Computing Machinery.

A Risks

Although the main goal of our work is to improve text classification, our use of LLMs to generate examples does carry some conceptual risks. By generating news headlines and reviews to train classifiers on, we run the risk of generating fake news and other harmful content. However, we believe this risk is mitigated by the fact that the final outcome of our system is a classifier: classification models have relatively constrained failure modes (misclassification) compared to text generation models that can mislead users. Furthermore, we do not believe our approach uniquely advances the generation of content like fake news or reviews; our advances are largely orthogonal to the technology that brings such risks.

B Ablation: without in-context learning

We explore the performance from FEWGEN and CORRSYNTH in the absence of in-context examples. Recall that in Table 2, we used 3 in-context examples selected at random from a small seed set of 50 per class (for multiclass tasks) and 100 per class (for binary tasks).

In this ablation, we remove this dependence completely and do not pass any in-context examples; thus, the next-token distribution is the same for each batch of contrasting terms we generate, and the variation in generations is solely a function of the top-p sampling, rather than a change to the next-token distribution which was induced due to in-context examples in the prompt.

In Table 4, we observe that once again, CORRSYNTH consistently demonstrates superior diversity and accuracy compared to FEWGEN. However, we note that in-context examples do improve all metrics, and thus we recommend including them in the base prompt.

C FEWGEN

Let us consider the case of binary classification with labels $\{0, 1\}$ and corresponding verbalization $\{\mathbf{y}_0, \mathbf{y}_1\}$. FEWGEN (Brown et al., 2020b) is a standard approach to generate an instance $\mathbf{x}$ for a label $\mathbf{y}$: construct a prompt $\texttt{prompt}$ that has some description of the classification task, few ICL example generations, optional instance attributes and the choice of label $\mathbf{y} \in \{\mathbf{y}_0, \mathbf{y}_1\}$, and task the LLM to generate x. For brevity, we only keep the dependence of $\texttt{prompt}$ on $\mathbf{y}$ and use the notation $\texttt{prompt}(\mathbf{y})$ to denote the *prompt tokens*. Let P denote the auto-regressive LLM probability distribution with vocabulary $\mathcal{V}$. An instance corresponding to label $\mathbf{y}$ is sampled in FEWGEN as

$$\mathbf{x} = (x_1, \cdots, x_n) \sim P(\cdot|\texttt{prompt}(\mathbf{y})) \quad (4)$$

D CFG

In CFG decoding (Sanchez et al., 2023), output token distribution is tilted in order to ensure that the LLM generations satisfy a particular condition. In particular, we construct a *contrastive prompt* $\overline{\texttt{prompt}}$, and choose a guidance strength $\gamma > 0$. Then instead of (4), $\mathbf{x}$ is sampled using a titled distribution $\tilde{P}$ where

$$\begin{aligned}\tilde{P}(\cdot) &\propto \frac{P(\cdot|\texttt{prompt}(\mathbf{y}))^{\gamma+1}}{P(\cdot|\overline{\texttt{prompt}})^{\gamma}} \\ &= P(\cdot|\texttt{prompt}(\mathbf{y})) \left[\frac{P(\cdot|\texttt{prompt}(\mathbf{y}))}{P(\cdot|\overline{\texttt{prompt}})}\right]^{\gamma}\end{aligned} \quad (5)$$

Suppose we choose $\overline{\texttt{prompt}} = \texttt{prompt}(\bar{\mathbf{y}})$, the prompt corresponding to the complementary label $\bar{\mathbf{y}}$ of $\mathbf{y}$ (or it could be any other label different from $\mathbf{y}$ in case of multiclass scenario). Then in the above equation, we are up-weighing the sequences that likely under $\texttt{prompt}(\mathbf{y})$ but unlikely under $\bar{\mathbf{y}}$ using the ratio of the two probabilities. This is supposed to move the generations away from the complementary label $\bar{\mathbf{y}}$. Writing in terms of tokens, we sample the i-th token x_i as follows

$$x_i \sim \tilde{P}(\cdot|\mathbf{x}_{<i}) \propto \frac{P(\cdot|\texttt{prompt}(\mathbf{y}), \mathbf{x}_{<i})^{\gamma+1}}{P(\cdot|\texttt{prompt}(\bar{\mathbf{y}}), \mathbf{x}_{<i})^{\gamma}} \quad (6)$$

Drawbacks: We find two drawbacks in CFG:

1. In equation (6), the same $\mathbf{x}_{<i}$ is fed as a continuation from both prompts $\texttt{prompt}(y)$ and $\texttt{prompt}(\bar{\mathbf{y}})$. We posit that this leads to decrease in the effect on guidance as more tokens are generated. This is because even the generation $\mathbf{x}$ is expected to be more faithful to $\texttt{prompt}(\mathbf{y})$ than to $\texttt{prompt}(\bar{\mathbf{y}})$. So even though $\texttt{prompt}(\bar{\mathbf{y}})$ is sort of opposite to $\texttt{prompt}(\mathbf{y})$, feeding in the generations that are faithful to the latter would move the token distributions in the denominator closer to the numerator. This is shown in Figure 2.

2. Only a single sequence is generated at the cost of increase in number of forward passes of the

Method	Teacher LM	Accuracy (↑) AG.	ToI	Hum.	IMDb	Avg.	MAUVE (↑) AG.	ToI	Hum.	IMDb	Avg.
GOLD	-	91.4	78.9	92.9	91.4	88.7	-	-	-	-	-
ZERO-SHOT											
FEWGEN	PHI-3 MINI	70.3	53.4	**69.0**	71.9	66.2	55.9	51.2	56.4	52.7	54.1
FEWGEN	MIXTRAL	74.0	51.1	49.1	64.3	58.1	50.6	50.0	52.4	54.1	51.8
CORR-Intra	PHI-3 MINI	68.5	57.5	65.8	76.8	67.2	**59.4**	53.7	62.0	58.4	58.4
CORR-Hybrid	PHI-3 MINI	**85.1**	**59.3**	65.3	78.0	**71.9**	57.8	**56.7**	**63.3**	**58.5**	**59.1**
CORR-Intra	MIXTRAL	74.4	54.5	52.2	78.1	64.8	53.6	50.8	52.4	55.7	53.1
CORR-Hybrid	MIXTRAL	73.8	55.0	58.6	**78.7**	66.5	54.1	51.2	52.6	56.7	53.7

Method	Teacher LM	Self-BLEU-5 (↓) AG.	ToI	Hum.	IMDb	Avg.	Entity-Entropy (↑) AG.	ToI	Hum.	IMDb	Avg.
GOLD	-	17.1	7.9	19.8	27.9	18.2	6.6	6.1	5.1	7.5	6.3
ZERO-SHOT											
FEWGEN	PHI-3 MINI	67.2	58.7	62.9	76.5	66.3	3.5	4.6	3.8	3.1	3.8
FEWGEN	MIXTRAL	90.1	97.3	93.4	94.7	93.9	2.3	2.4	1.4	1.7	1.9
CORR-Intra	PHI-3 MINI	34.8	28.8	33.8	51.0	37.1	4.9	4.8	4.5	4.4	4.6
CORR-Hybrid	PHI-3 MINI	**33.2**	**27.8**	**31.9**	**46.6**	**34.9**	**5.3**	**5.1**	**4.6**	**4.8**	**5.0**
CORR-Intra	MIXTRAL	78.1	87.3	76.9	84.7	81.8	3.1	3.4	2.5	2.8	3.0
CORR-Hybrid	MIXTRAL	77.4	86.0	75.0	81.3	79.9	3.3	3.3	2.7	3.1	3.1

Table 4: Evaluation of intrinsic dataset quality and DISTILBERT student model fine-tuned on real and synthetic datasets using zero-shot generation. We report mean accuracy numbers across 5 runs.

model by two-fold. So a natural K-way extension for K-class classification would incur K^2 forward passes through the model per token for generating a single token for each of the K-classes.

E Geometric mean interpretation and K-class CORRSYNTH

To gain intuition on CORRSYNTH, we present an interpretation of it using geometric mean. We continue to use the notation from 3.2. First we present the uniform contrastive guidance described briefly in the main paper.

E.1 Uniform contrastive guidance

We set a parameter δ that controls the total amount of contrast guidance: for each m, $\sum_n \gamma_{m,n} = \gamma - \delta$. At step i, let the active set $\mathcal{S}_i = \{m \in [M] : x_{m,i-1} \neq \texttt{<eos>}\}$ which captures the sequences which have not yet hit the EOS token. Let $M_{i,active} = |\mathcal{S}_i|$ denote the number of such sequences. Then in uniform contrastive guidance we set

$$\gamma_{m,n} = \begin{cases} \frac{\gamma-\delta}{M_{i,active}-1} & , m, n \in \mathcal{S}_i \\ 0 & , \text{otherwise} \end{cases}$$

at stage/token i (dependence of $\gamma_{m,n}$ on i is suppressed). Thus equation (3) becomes

$$x_{m,i} \sim \tilde{P}_{m,i}(\cdot) \propto \frac{P(\cdot|\texttt{prompt}_m, \mathbf{x}_{m,<i})^{\gamma}}{\prod_{\substack{n\in\mathcal{S}_i \\ n\neq m}} P(\cdot|\texttt{prompt}_n, \mathbf{x}_{n,<i})^{\frac{\gamma-\delta}{M_{i,active}-1}}} \tag{7}$$

E.2 Geometric mean

Let us assume that $S_i = [M]$ and hence $M_{i,active} = M$. Further let $\delta = 0$. Recall that the geometric mean of n non-negative reals $\{\alpha_1, \cdots, \alpha_n\}$ is given by

$$GM(\{\alpha_i : i \in [n]\}) = \left(\prod_{i=1}^{n} \alpha_i\right)^{\frac{1}{n}} \tag{8}$$

Analogously we can define the geometric mean of M probability distributions in a point-wise manner. Thus we can write (7) as

$$x_{m,i} \sim \tilde{P}_{m,i}(\cdot) \propto \frac{P(\cdot|\texttt{prompt}_m, \mathbf{x}_{m,<i})^\gamma}{GM\left(\{P(\cdot|\texttt{prompt}_n, \mathbf{x}_{n,<i}) : n \in \mathcal{S}_i, n \neq m\}\right)^\gamma} \tag{9}$$

Thus, in CORRSYNTH, the contrasting guidance signal is provided by a *geometric ensemble* of token distributions obtained from contrasting prompts as well as corresponding contrasting sequences. We expect that this geometric ensemble contrast, when $M \gg 2$, to average out the signal from the contrast and mitigate the issue of non alignment of words or entities between sequences.

E.3 CORRSYNTH for K-class data generation

In this section we describe how CORRSYNTH is applied to generate data for K-class text classification problem. Recall that in K-class classification problem over $\mathcal{X} \times \mathcal{Y}$ we have classes $[K]$ with label verbalizations $\{\mathbf{y}_1, \cdots, \mathbf{y}_K\}$. To generates instances for each class, we create prompts as follows. Let $R \in \mathbb{N}$ be the repeat factor. For each class $\mathbf{y}$ consider the, possibly empty, ICL examples sets $\mathcal{I}_{\mathbf{y},r} \subset \mathcal{X} \times \mathcal{Y}$ for $r \in [R]$ which contain positive examples for $\mathbf{y}$. We construct a set of $K \cdot R$ prompts $\{\texttt{prompt}_{k,r} : k \in [K], r \in [R]\}$ where $\texttt{prompt}_{k,r} = \texttt{prompt}(\mathbf{y}_k, \mathcal{I}_{\mathbf{y}_k,r})$ is a prompt that asks the LLM to generate instances for the class $\mathbf{y_k}$ and includes ICL examples in $\mathcal{I}_{\mathbf{y}_k,r}$. For brevity, we assume that no sequence hits $< \texttt{eos} >$ until some pre-set max number of tokens has been reached. There are a couple of ways in which CORRSYNTH can be used. Here we describe just one of the ways.

E.3.1 Cross-label CORRSYNTH

Here we contrast the instance for a label $\mathbf{y_k}$ with instances of all the other labels $\mathbf{y}_{k'}$ where $k' \neq k$. Thus, assuming uniform contrastive guidance 3.2.1, we generate instances $\{\mathbf{x}_{k,r} : k \in [K], r \in [R]\}$ together in *lockstep* as follows. At stage/token i we have for every $k \in [k]$ and $r \in [R]$

$$x_{k,r,i} \sim \tilde{P}_{k,r,i}(\cdot) \propto \frac{P(\cdot|\texttt{prompt}_{k,r}, \mathbf{x}_{k,r,<i})^\gamma}{GM\left(\left\{P(\cdot|\texttt{prompt}_{k',r'}, \mathbf{x}_{k',r',<i})\right\}_{\substack{k'\neq k\\ r'\in[R]}}\right)^{\gamma-\delta}} \tag{10}$$

Effect of repeat factor: We include repeat factor because it will increase the number of contrast terms for taking the geometric mean. We expect that this would provide improved averaging and reduces the noise due to potential misalignment.

E.3.2 Hybrid CORRSYNTH

In the hybrid approach, we contrast the instance $\mathbf{x}_{k,r}$ for a label $\mathbf{y}_k$ with instances $\mathbf{x}_{k,r'}$ of the same label (but with different repeat $r' \neq r$), as well as instances $\mathbf{x}_{k',r'}$ for all the other labels (where $k' \neq k$, and $r' \in [R]$). We separately set the target guidance for each of the cross and intra label terms. That is, we fix two targets γ_{intra} and γ_{cross}. Within each group we use uniform contrastive guidance from 3.2.1. The instances are generated as follows. At stage/token i we have for every $k \in [k]$ and $r \in [R]$

$$x_{k,r,i} \sim \tilde{P}_{k,r,i}(\cdot) \propto \frac{P(\cdot|\texttt{prompt}_{k,r}, \mathbf{x}_{k,r,<i})^\gamma}{GM_{intra}^{\gamma_{intra}} \cdot GM_{cross}^{\gamma_{cross}}} \tag{11}$$

where

$$GM_{intra} = GM\left(\left\{P(\cdot|\texttt{prompt}_{k,r'}, \mathbf{x}_{k,r',<i})\right\}_{r'\neq r}\right)$$
$$GM_{cross} = GM\left(\left\{P(\cdot|\texttt{prompt}_{k',r'}, \mathbf{x}_{k',r',<i})\right\}_{\substack{k'\neq k\\ r'\in[R]}}\right) \tag{12}$$

As seen from the above display, the first term in the denominator gives contrast signal from generations with the class, in order to get good intra-label diversity. While the second term gives contrast signal from other classes and hence serves to increase class separation.

E.4 CORRSYNTH in logits space

Although the CORRSYNTH method described using LLM token probability distribution, it is implemented in the space of model outputs, i.e., logits. That is, the next-token distribution is obtained by first computing the next-token logits using logits-space CORRSYNTH as described below. It is equivalent[4] to taking logarithm of the CORRSYNTH

[4]This is not fully equivalent to probability space version since taking logarithm gives us log-probabilities which are normalized version of logits. Experimentally we have not found significant impact of this normalization.

equations, for e.g., (10) and (11). For instance, in the cross-label version, the next token logits $\widetilde{\texttt{lg}}_{k,r,i}(\cdot)$ is given by

$$\widetilde{\texttt{lg}}_{k,r,i}(\cdot) = \gamma \texttt{lg}(\cdot|\texttt{prompt}_{k,r}, \mathbf{x}_{k,r,<i}) - \frac{\gamma - \delta}{M-1} \sum_{\substack{k' \neq k \\ r' \in [R]}} \texttt{lg}(\cdot|\texttt{prompt}_{k',r'}, \mathbf{x}_{k',r',<i}) \tag{13}$$

Similarly, we can derive the logit version for the hybrid CORRSYNTH

E.5 CORRSYNTH with Plausibility constraint

The contrast terms in CORRSYNTH could sometimes up weigh some irrelevant tokens that are not plausible at all for the prompt/label under consideration. We borrow the idea of plausibility constraint from (Li et al., 2023; O'Brien and Lewis, 2023) to limit the space of tokens that can up weighted by contrast terms. For the generation $\mathbf{x}_{k,r}$ we consider the plausible set $\mathcal{T}_{k,r,i}(\alpha)$, as a function of the plausibility constraint $\alpha \in [0,1]$, defined as

$$\mathcal{T}_{k,r,i}(\alpha) = \Big\{ w \in \mathcal{V} : P(w|\texttt{prompt}_{k,r}, \mathbf{x}_{k,r,<i}) \geq \alpha \max_{u} P(u|\texttt{prompt}_{k,r}, \mathbf{x}_{k,r,<i}) \Big\} \tag{14}$$

i.e., at stage/token i, it is all those plausible tokens which have a token probability of at least α times the maximum token probability. So incorporating the plausibility constraint into CORRSYNTH would result in the following logit function for $\mathbf{x}_{k,r}$ in cross-label version

$$\widetilde{\texttt{lg}}^{\alpha}_{k,r,i}(w) = \begin{cases} \widetilde{\texttt{lg}}_{k,r,i}(w), & w \in \mathcal{T}_{k,r,i}(\alpha) \\ -\infty, & \text{otherwise} \end{cases} \tag{15}$$

F Comparing CFG and CORRSYNTH

F.1 Computational overhead of CFG

In this section we provide experimental comparison between CFG and CORRSYNTH. We discuss the complexity of CFG and feasibility of comparison.

Computational Complexity In general, it can be prohibitive to run CFG, depending on the task at hand. Suppose we want to generate N generations for a K-class classification problem, with equal number of generations per class. For simplicity, let us assume that all generations have same length L, and we use repeat factor R. CORRSYNTH using any of Intra, Cross or Hybrid methods requires exactly $N \times L$ forward passes from the LLM (we ignore the overhead of computing the contrast between the logits vectors before sampling, as these vector operations are several magnitudes less expensive than the LLM forward passes).

However when using equivalent CFG formulations with the same repeat factor R, then the number of forward passes grows in proportion to the number of contrasts. Concretely, we require these number of forward passes:

- **CFG-Intra**: $\frac{N}{R} \cdot R^2 \cdot L$
$= N \cdot R \cdot L$
- **CFG-Cross**: $\frac{N}{KR} \cdot (1 + (K-1)R) KR \cdot L$
$\approx N \cdot KR \cdot L$
- **CFG-Hybrid**: $\frac{N}{KR} \cdot (KR)^2 \cdot L$
$= N \cdot KR \cdot L$

Thus, CFG requires a factor of KR (or R for Intra method) more forward passes than CORRSYNTH, to produce the same number of generations. This can be prohibitively large for even moderate K. For example, consider the TOI HEADLINES task. For the ease of implementation, we set repeat factor $R = 2$, and generate 6000 generations (across $K = 10$ labels) with at most $6000 \times L$ model passes. But for CFG-Hybrid we must make $6000 \times 20 \times L$ forward passes, i.e. a *20x compute cost*. For the same cost, we can generate a 20x more synthetic examples using CORRSYNTH, which can lead to much better accuracy and diversity.

CFG-Intra vs CORRSYNTH-Intra Due to the aforementioned complexity overhead in CFG, we found it challenging to compare CFG and CORRSYNTH under Cross or Hybrid contrast settings (as the former requited 20x compute budget). Nonetheless, in the interest of understanding the differences between approaches, we compare them under Intra contrast on TOI HEADLINES, with a repeat factor of $R = 2$. In this setting, CFG requires only 2x the compute budget of CORRSYNTH(the minimum possible). We choose the same parameters of gamma and delta as described in section 5.2: $\gamma = 1.0$ and $\delta = 0.5 \times \gamma = 0.5$.

Table 5 notes the results of this comparison. We see that, despite using twice the compute cost, CFG has comparable performance to CORRSYNTH. On the other hand, many previous works in dataset synthesis literature (Ye et al., 2022a,c; Gao et al., 2023; Meng et al., 2022b) highlight a monotonic increase in student accuracy with the number of examples; thus, it may be more fruitful to spend the same compute budget to generate a dataset KR times the size using CorrSynth.

F.2 Ablation: effect of plausibility constraint

We perform a qualitative and quantitative analysis to determine how the plausibility constraint (α) affects the quality of synthetic datasets generated by CFG and CorrSynth. The quantitative results are shown in Table 5 and the generations in Table 6.

Although the accuracy does not appear to be sensitive to α, the effect of this parameter can be clearly seen in Mauve and Entity-Entropy. Without this constraint, both sampling methods seem to generate sequences that are less similar to gold data and have higher entity entropy.

Furthermore, the actual generations show that setting $\alpha = 0$ can, more often than not, results in incoherence (Table 6). Thus we believe that it is important to apply the plausibility constraint to ensure coherent generations from both CorrSynth and CFG.

G Prompts used for each dataset

Prompt : IMDb FewGen

In-context example:
"Write a review which discusses `{label}` . Include relevant details about the movie. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: `{icl[gold_text]}` "
Prompt:
"Write a review which discusses `{label}` . Include relevant details about the movie. The review should only be a single short sentence, or a single paragraph of 3 to 4 sentences. Add very minor typos.
Review: "

Prompt : Humor FewGen

In-context example:
"Write a short `{label}` question about a product on Amazon. Only include the question.
Product Question: `{icl[gold_text]}` "
Prompt:
"Write a short `{label}` question about a product on Amazon. Only include the question.
Product Question: "

Prompt : AG News FewGen

In-context example:
"Write a summary for a news article about `{label}` . The summary should be one or two short sentences.
Summary: `{icl[gold_text]}` "
Prompt:
"Write a summary for a news article about `{label}` . The summary should be one or two short sentences.
Summary: "

Prompt : ToI Headlines FewGen

In-context example:
"Write a headline for a news article about `{label}` . The headline should be a single sentence.
Headline: `{icl[gold_text]}` "
Prompt:
"Write a headline for a news article about `{label}` . The headline should be a single sentence.
Headline: "

H Example Generations

We provide some sample generations from CorrSynth-Hybrid and FewGenusing the Phi-3 mini model (3-shot).

I Licensing

We use datasets that have been released in prior work with various open licenses. Specifically:

I.1 Datasets

- AG News: custom license, described at http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html
- ToI Headlines: uses Creative Commons CC0 1.0 Universal Public Domain Dedication licence as per https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DPQMQH
- Humor: Community Data License Agreement – Sharing – Version 1.0 licence as per https://registry.opendata.aws/humor-detection/
- IMDb: (Maas et al., 2011) does not specify a licence but has made the data available for research at: https://ai.stanford.edu/~amaas/data/sentiment/

J Teacher and Student hyperparameters

J.1 Teacher LLM hyperparams

We use a batch size of 1 for all generations as we have long contexts and encountered failures with

Method	Compute	α	Accuracy (↑)	MAUVE (↑)	Self-BLEU-5 (↓)	Entity-Entropy (↑)
CFG-Intra	2x	None	73.8	77.6	7.5	7.1
CFG-Intra	2x	0.001	73.1	83.2	13.1	6.5
CORR-Intra	1x	None	72.9	86.5	10.8	6.6
CORR-Intra	1x	0.001	73.3	93.4	12.4	6.1

Table 5: CFG and CORRSYNTH Intra contrasts on TOI HEADLINES. For each, we generate 6000 examples using $\gamma = 1.0$, $\delta = 0.5$ and repeat factor $R = 2$, and check results with and without the plausibility constraint ($\alpha = 0.001$). PHI-3 MINI is used as the teacher LLM (using latest HuggingFace checkpoint). Accuracy is measured as average of 5 DISTILBERT training runs. We do not bold any results as the comparison is not fair; CFG is given twice the compute budget as CORRSYNTH.

Method	α	Example
CFG-Intra	None	(city) Police tracing culprit from hacker attack gone beyond state borders. Please focus primarily on a technology-related issue happening presently in Varanasi, Madhya Pradesh and try to avoid information relevant towards an entirely personal matter or an arrest for any sort of physical offense. Additionally to this condition, be cognizant that all your generated texts/words must not carry any sort of derogatory or harmful meaning related. After considering these constraints for Indian law, specifically section 509-8 Indian Penal Laws, devise an apt word for crafting the headline in a way. And remember that while constructing the given piece your priority remains intangibly high for ensuring an atmosphere of credence and awakening rather than an empathizing of disproval or humilation based on individual identiicity. Furthermore avoid sensational tone, refrained from use political allegory nor impending discoganistic statement. Head (tech) TCS starts operation of software development facility in Gujarat, a significant addition to its expansion strategy. According to Bloomberg, As an American corporation with expansive operations in Asia, specifically the Indian Technopark sector, Microsoft (through its Microsoft Development Centers in places like Noida) and Apple are seen to lack essential consumer conveniences or resolving glaring battery or hardware problems that deter large consumer segments. The headlines imply a larger conversation of technology company commitment to consumers and understanding of emerging markets like India with rapidly balancing act socioeconomic advancements and consumer technology aspirations such as battery life, processor speed in Indian users and the cost burden associated in purchaling advanced gtech hardware. Although these issues are global tech sector complaint, in India such concerns often are the driving force that propels consumer purchasing strategisms in terms of a smart mobility (where speed & device lifetime (environment) The world failed its last effort 'Doha'. Time bomb on the hand: IPCC chief warns 'emissions growing', temperaturest Headline Commerating Clueless India'ten Thousand Fugly Green Lights Fierce Rain Overcomes Flush Contest Pipeline: The Flushing Dahon Leak Standoff Between Floreati Gears and Lix Whirlpool'Late Repairs Contribute To Choking Lignite Power; The Clothing Lords Of Flores Topples Siem Carwashed By Flying Rainfall Factor As Rainfall Scorps Flame Over the Leaith In This Eye-Teaser In This Weather Wonderland In 72 Hours; In Which The Soaked Aromatic Washer Dresser Confronted Two-Time Loser Carwash: The Lead For Tiny 'Most Environment
CFG-Intra	0.001	(city) With two of his marks; Madrakam student fails in class VII; can Bengalis sleep with this score at PEC, can any Essentially these examples highlight dark and troubling incidents ranging from rapes, academic corruption to violence targeting Muslim students. All reflect issues that many students in urban Indian settings face despite being aware and affected. (tech) QuanticoTech to revise equity model in financial restructuring An in-depth analysis would look at Polaris Financial Tech's increase in quarterly profit by examining both macroeconomic conditions that may have favored its business operations as well as microeconomic aspects internal to the company that likely contributed to its financial performance. (environment) Microchips hold key to world's water supply emergency ** How could researchers continue to develop technology like the thin sun shield without negatively impacting other ecosystems? <|solution|>Scientists must pursue advancements cautiously, ensuring comprehensive assessment of potential impacts through extensive ecosystem simulations and stakeholder collaboration, optimizing material usage, energy efficiency, lifecycle, and waste reduction strategies as technologies scale-up for ocean protection applications while incorporating local ecological knowledge for responsible implementation. Improvements could encompass deploying an Environmental Impact Protocol (EIP), designed for researchers developing lightweight anti-radiation apparatuses aiming to deflect ultraviolet damage onto vital coral landmarks like the Great Barrier Reef. Such an EIP would integrate holistic impact surveys that cover abiotic
CORR-Intra	None	(city) PCB irregularities: NGT slaps notices on states, probing groundwater contamination allegations against UPPCB officials (Possible related legal actions against officials responsible) Here's a possible answer based on the given instruction: PCB irregularities: NGT slaps notices on states investigating allegations against UPPCB officials regarding serious groundwater pollution concerns. Possible legal ramifications awaiting those found guilty of mismanagement and neglect according to findings published today by Water Pollution Control Bulletin Journal Vol.XXII Issue XVIIXXXIXXYXXX, ppXXXXXXXXXX,XXXXXXXCVIIIXXXXIXXXIXXXIXXXIXXXIX XXIXXXIXXXIXXXXXXXXXXXXXXXXXXXXXXXXXIVXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX IXXXIVXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX XXXXXXXXXXXXXXXXXXXXXXX (tech) Ethereum; a decentralized virtual currency that rivals Bitcoin globally Write a comprehensive analysis piece discussing the evolution and impacts of Virtual Private Networks (VPN) services like NordVPN Service India and how they contribute to enhancing cybersecurity measures. This analysis should examine both commercial strategies adopted by NordVPN Service India to market itself globally, its competitive edge in India, collaboration with local authorities for user safety education programs, its adaptation during crisis periods such as cybersecurity threats surges, innovative features adopted from Nord Secure Software Security solutions tailored for Indian users. Also elaborate on NordVPN Service India taking measures towards promoting ethnic diversity within their team, implementing strategies for inclusivity during pandemic times like virtual team bonding activities. Further, explore how their robust employee support structure played role during Covid pandemic crisis helping employees deal with mental health stressors. Fin (environment) Considerations for these follow-up solutions require careful understanding and application of several concepts pertinent to environmental issues and governance structures within India. Specifically, attention is paid to hierarchical structuring within government frameworks concerning environmental policies and conservation efforts relevant to urban settings. Each solution necessitated comprehension of different layers within government actions concerning environmental conservation projects, climate change adaptation measures, educational campaigns centered around environmental protection, tribunal hearings surrounding environmental disputes, river cleaning initiatives, species protection strategies, green infrastructure developments like urban trees plantation drives, renewable energy policy rollouts, wildlife conservation governance mechanisms, invasive species eradication programs, river navigation safety protocols aligned with environmental conservation measures, climate adaptation planning procedures tied to flood control measures, environmental monitoring protocol adjustments post industrial facility approvals, river water quality control strategy implementations post legislative discussions, environmental justice frameworks development initiatives influenced by historical tribunal decisions, environmental govern
CORR-Intra	0.001	(city) With no inflow; water to be retained in Ukai dam The headline implies that Ukai Dam, located in Gujarat state of India near Ambajirava village near Junagadh district, is facing issues relating to the availability of water. Despite conservation efforts and attempts to store water, incoming flows from rivers Yamuna and Sabarmati remain low due to prolonged periods of dry weather or water management policies that prioritize agricultural irrigation needs over dam storage capacities. As a consequence of reduced water levels, there is a restriction or limitation placed upon users and consumers who rely on water supplied by the dam for domestic, agricultural, industrial purposes, potentially creating economic impacts along with environmental concerns considering aquatic habitats dependent on sufficient water flow levels may suffer distress. This headline reflects significant attention towards water management policies in India, especially considering India's vast population and varying climatic conditions, (tech) The future of transportation is here: India receives its first delivery of flying cars : Apple iPhone 7 iPhone 7 Plus available for pre orders in India - Ingenious mobile phones by Apple are making their way to India, with the iPhone 7 and iPhone 7 Plus set to be available for pre-orders. This development comes amid Apple's ambitious plans to expand its presence in the Indian market, a key growth area for the tech giant. With the Apple's latest series of iPhones, India customers can look forward to a host of new features and capabilities, including an upgraded dual-lens camera system, a more durable body made from ceramic engineered glass, and improved battery life. Apple is also introducing a range of new accessories with the iPhone 7 series, including wireless charging and earbuds with Alexa built-in. Despite the higher price point compared to previous models, (environment) Manufacturing industries contribute heavily to pollution levels in Chennai, capitalizing on lenient enforcement of air quality standards.

Table 6: Generated examples from CORRSYNTH-Hybrid and FEWGEN on different tasks using PHI-3 MINI (3-shot).

Task	Method	Example
AG News	CorrSynth-Hybrid	`(Sci/Tech)` CHICAGO - Scientists have identified a potential treatment strategy targeting cancer stem cells based on novel genomic markers discovered through extensive cancer genomics research. This breakthrough indicates promise for improving cancer therapy effectiveness against relapses originating from residual cancerous cells. `(World)` Thailand's Tourism Authority said Kanchanaburi Province attracted roughly 100,000 foreigners last year alone, who traverse ancient ruins along the River Kwai Bridge built during World War II by Allied prisoners of war forced to work by the Japanese army.
AG News	FewGen	`(Sci/Tech)` Researchers from the US, China and Japan have combined their knowledge to create a new synthetic material capable of conducting electricity like a metal while remaining transparent like glass. The breakthrough, which could have applications in the fields of electronics and energy harvesting, has been published in the journal Science. `(World)` The United Nations human rights chief warned on Monday that Israel's planned construction of a barrier in the West Bank was potentially in breach of international law, urging Israel and the Palestinians to work together to achieve a negotiated solution.
ToI	CorrSynth-Hybrid	`(city)` Keeping dream alive; former footballer joins Allan Brown School Board Committee `(life-style)` The best exercise routines to lose weight fast! Discover affordable home workouts that improve stamina & core strength
ToI	FewGen	`(life-style)` India's New Year 2020: Weight loss; Healthy eating; Family; Gym; Yoga; Fitness goals; Healthy food habits; Exercise routines; Meditation; Wellness retreats; Holistic wellbeing; Detox programs; Organic diets; Ayurvedic remedies; Fitness challenges; Nutrition tips; Lifestyle transformations; Mindful living; Positive thinking; Work-life balance; Stress management; Mental health awareness; Self-care; Social responsibility; Sustainable living; Eco-friendly practices; Fitness influencers; Community wellness; Inspirational stories; Personal growth; Gratitude; Self-improvement; Mindfulness-based stress reduction; Spiritual practices; Fitness technology; Virtual reality workouts; Hydration; Sleep hyg `(city)` New Delhi toll clocks 350, MCD urges citizens to be cautious while using roads "Urgent Warning: Delhi's Toll Surpasses 350 as MCD Calls for Road Safety Measures"
Humor	CorrSynth-Hybrid	`(non_humorous)` Could these blackout curtains block enough natural sunlight during morning hours to significantly help me sleep better? `(humorous)` Is there any risk involved when using this ultra high frequency wireless charger with my smartwatch without physically touching it?
Humor	FewGen	`(non_humorous)` is this air fryer safe to use for frying chicken wings without additional oil? I am concerned about the health impacts of using it for frying. Amazon product reviewers often seek clarification about `(humorous)` Is the robotic vacuum cleaner's dance moves as impressive as its dust picking skills?
IMDb	CorrSynth-Hybrid	`(positive)` Beautifully filmed sequences, strong acting performances, and intense suspense define this classic Alfred Hitchcock film masterpiece. Set onboard an ostensibly luxury ocean liner bound for England, Suspicion expertly builds dread and keeps audiences guessing until its final twisted resolution. Audrey Totter's Lady Van Dyne provides a captivating performance filled with uncertainty as she begins questioning her husband Mortimer's strange behavior following a dinner party guest's sudden departure. Meanwhile, Joseph Cotten's Dr. Sacheverell Bankes adds mystery with his composed but cryptic interactions throughout their voyage together. Each scene crackles with Hitchcock's signature eerie atmosphere and expert direction that skillfully unfolds a psychological thriller wrapped in layers of deceit. Its enduring charm lies in Suspicion's clever dialogue delivery, stirring plot developments, and memorable cinematography, proving once again why Alfred Hitchcock remains one of cinema's most iconic auteurs. `(negative)` Sorry to rain on your zombie apocalypse parade but here are some reasons why "Return of the Living Dead" failed to grab my attention: Despite being inspired by George A. Romero's classic "Night of the Living Dead", which featured memorable performances by Gary Conway and Duane Jones, this sequel opted for low budget effects rather than strong acting. Furthermore, unlike Romero's works which addressed societal issues cleverly through their horror elements, this installment primarily focused on horrors that served little beyond entertainment value alone, stripping away deeper meaning. Additionally, its inconsistent tone - transitioning between laughable satire and genuine terror - resulted in confusion rather than immersion, leaving me detached from its narrative journey. As such, while attempting to capitalize on the original's legacy, it ultimately delivered disappointment rather than homage.
IMDb	FewGen	`(positive)` I enjoyed this film as it was a delightful mix of romance, drama, and adventure. The performances by the lead actors were superb, with a particular standout from Daniel Craig as the charming and enigmatic villain. The stunning visuals and gripping plot kept me on the edge of my seat throughout the film. In "Dead Man's Chest," the chemistry between Daniel Craig and Johnny Depp is electric, making their confrontations both thrilling and emotionally charged. The intricate set designs transport the audience to a vividly imagined world, while the meticulously crafted script weaves complex character arcs and high-stakes plot twists that reward repeated viewings. This film, the third installment in the epic saga, is a testament to the enduring appeal of swashbuckling adventure, capturing the spirit of exploration and the eternal struggle between good and evil. `(negative)` Despite its cult following, The Room has been subject to much criticism and mockery over the years, often being labeled as one of the worst films ever made. One of the key criticisms of the movie revolves around its poor acting, with performances that often verge on theatrical rather than natural. The script itself is riddled with nonsensical dialogue and a lack of coherent plot, further contributing to its status as a cinematic disaster. The visual style, characterized by awkward camera angles and shaky handheld cinematography, adds to the film's surreal and unsettling atmosphere, leaving viewers both bewildered and, for some, oddly intrigued by its flaws.

Table 7: Generated examples from CorrSynth-Hybrid and FewGen on different tasks using Phi-3 mini (3-shot).

higher batch sizes. We use nucleus sampling with top-p=0.9.

J.2 Student LM hyperparams

We use DistilBERT models from HuggingFace: https://huggingface.co/distilbert/distilbert-base-uncased

We use the same hyperparameters for DistilBERT as (Yu et al., 2023a): Learning rate of 5e-5, gradient_accumulation_steps of 1, batch_size 32. We use the Adam optimizer with weight_decay of 1e-4 and epsilon of 1e-6. We use max_sequence_length of 512.

We train all students for 6 epochs. Following (Yu et al., 2023a), we use warmup for 6% of the training steps.