Create! Don't Repeat: A Paradigm Shift in Multi-Label Augmentation through Label Creative Generation

Letian Wang, Xianggen Liu, Jiancheng Lv[*]
College of Computer Science, Sichuan University, China
Engineering Research Center of Machine Learning and Industry Intelligence, China
letianwang@stu.scu.edu.cn, {liuxianggen, lvjiancheng}@scu.edu.cn

Abstract

We propose Label Creative Generation (LCG), a new paradigm in multi-label data augmentation. Beyond repeating data points with fixed labels, LCG creates new data by exploring innovative label combinations. Within LCG, we introduce Tail-Driven Conditional Augmentation (TDCA), combining tail-driven label sampling and label-conditioned text generation for balanced, consistent data augmentation. Our approach has demonstrated a **100.21%** increase in PSP@1 across three datasets, successfully mitigating the long-tail effect in MLTC and markedly enhancing model performance.

1 Introduction

Multi-label Text Classification (MLTC), prevalent in fields such as sentiment analysis and recommendation systems, grapples with the dual challenges of an extensive label space and a pronounced long-tail distribution. This imbalance is exemplified in Wiki10-31K dataset, where a mere **1.5%** of labels have more than **100** training instances, leaving the vast majority with a scarcity of training data. Table 1 reveals that while numerous studies claim to have addressed or alleviated the long-tail issue, advancements in recent years, particularly on PSP@k, have been gradual. Relying solely on advancements in neural architectures appears to be ineffective over time. There is a need for data augmentation.

Method	Source	PSP@1	PSP@3	PSP@5
PfastreXML	Jain et al. (2016)	19.02	18.34	18.43
XML-CNN	Liu et al. (2017)	9.39	10.00	10.20
AttentionXML	You et al. (2019)	15.57	16.80	17.82
LightXML	Jiang et al. (2021)	16.00	16.99	18.97
Cbolt	Ge et al. (2022)	12.00	13.50	15.00
XRR	Xiong et al. (2023)	11.77	16.48	21.07
TDCA	**ours**	**46.31**	**39.02**	**36.92**

Table 1: Comparison of MLTC methods on Wiki10-31K using $PSP@k$, a widely used metric that assigns higher weights to tail labels for a more balanced evaluation.

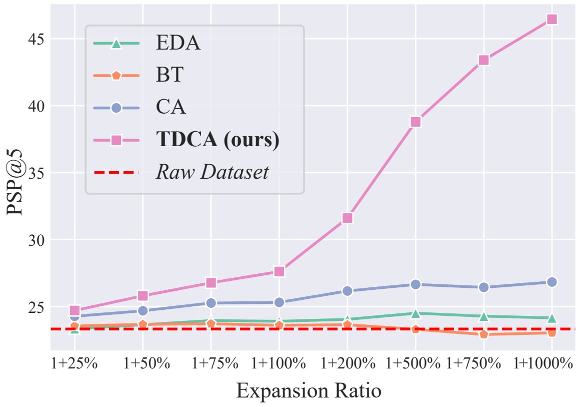


Figure 1: PSP@5 for various expansion ratios on MeSH-12K dataset, comparing four data augmentation methods: Easy Data Augmentation (EDA), Back-Translation (BT), Conditional Augmentation (CA), and TDCA. The Expansion Ratio (E/R) quantifies augmentation: 1 indicates no augmentation and $1 + x\%$ shows $x\%$ augmentation over the original dataset.

1.1 Inadequacies of Current DA Methods

Current data augmentation (DA) methods in text classification have shown promise, yet their effectiveness in MLTC remains limited. Prevailing DA strategies typically focus on employing various approaches to replicate data while maintaining labels unchanged. These methods fall into two distinct categories: paraphrase-based augmentation and conditional augmentation (CA). The former includes model-free methods like easy data augmentation (Wei and Zou, 2019), as well as model-required techniques such as back-translation (Sennrich et al., 2016; Edunov et al., 2018). CA employs conditional generation (often label-conditioned) for data augmentation to synthesize texts that are more controlled yet diverse (Li et al., 2020; Liu et al., 2020). Such strategies offered benefits in binary or multi-class text classification, often regarded as a means to bolster model robustness (Bayer et al., 2023). However, as illustrated in Figure 1, conventional approaches like EDA and BT are seldom effective and might even

[*] Corresponding Author.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 855–869
June 16-21, 2024 ©2024 Association for Computational Linguistics

impede model performance, particularly at greater data expansion scales. CA shows improved performance but remains modest.

Intrinsically marked by the existence of multiple labels per instance, MLTC poses distinct challenges to conventional data augmentation methods. Paraphrasing-based approaches, which solely alter the text while maintaining its original intent and keeping the labels unchanged, fail to enhance performance by a large margin. Moreover, the noise introduced in the paraphrasing process can adversely affect model performance, a negative impact that becomes more pronounced with an increasing proportion of augmented data.

1.2 LLMs in MLTC

The recent validation of scaling laws in large language models (LLMs) like GPT (Brown et al., 2020; Bubeck et al., 2023) and LLaMa (Touvron et al., 2023a,b) has revolutionized previously infeasible tasks and advanced tasks that had plateaued. Research exploring LLMs in multi-label text classification has been emerging. For example, Kocon et al. (2023) evaluated ChatGPT in text classification, observing its lag behind traditional SOTA models, with the gap widening in complex classification scenarios. However, its performance in emotion recognition was noteworthy. Loukas et al. (2023) corroborated this, highlighting ChatGPT's potential in few-shot and zero-shot classification. This is in line with the training goal of LLMs, which involves predicting the next token from preceding ones.

Post-training on large-scale general corpora, their formidable understanding and common sense skills lend them efficacy in text generation tasks and common classification tasks with fewer categories (Ray, 2023). Nonetheless, they grapple with issues like hallucinations and prompt sensitivity, underperforming in knowledge-intensive tasks like large-scale information retrieval and fine-grained text classification (Li et al., 2023).

Given the extremely large number of labels, employing LLMs for MLTC directly is impractical. An alternative is leveraging LLMs for data augmentation, as explored in some recent works. AugGPT (Dai et al., 2023), for instance, continued with the conventional approach, employing ChatGPT for paraphrasing texts while preserving the original labels, yielding marginal improvements. Van Nooten and Daelemans (2023) took a more straightforward route. They provided ChatGPT with nine distinct labels along with the corresponding examples to generate additional text-labels pairs for tweet classification.

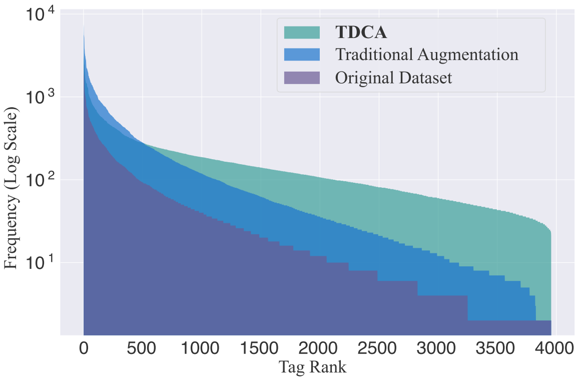


Figure 2: Tag frequency distribution post data augmentation. Compared to traditional data augmentation methods (EDA, BT, and CA), TDCA demonstrates more effective mitigation of the long-tail effect.

1.3 Create! Don't Repeat

The untapped potential of LLMs opens new avenues for conditional data generation to liberate us from the limitations of simply repeating existing data points, enabling the synthesis of entirely new ones. Hence, we introduce a new multi-label data augmentation paradigm, Label Creative Generation (LCG), and within LCG, we propose Tail-Driven Conditional Augmentation (TDCA). TDCA comprises (1) a Metropolis-Hastings algorithm based tail-driven label sampling for crafting more balanced label combinations with consideration of label correlations; and (2) a contrastive label-conditioned generation approach, which fine-tunes LLMs to generate texts that not only accurately reflect each label in the sampled label combinations but also emulate the style of the original dataset.

TDCA enables the transference of extensive correlations from head labels to tail labels through LLM. Our experimental results demonstrate a substantial alleviation of the long-tail effect, evidenced by a **100.21%** average enhancement in PSP@1 across Eurlex-4K, MeSH-12K, and Wiki10-30K datasets. Remarkably, with increasing expansion ratios, we observe an ascending trend in PSP@k, devoid of the detrimental noise impact commonly associated with traditional data augmentation methods, a finding corroborated by t-SNE visualizations in Figure 5. Ablation studies affirm the efficacy and validity of TCDA.

Our contributions are threefold:

1. We introduce a new paradigm in multi-label data augmentation, Label Creative Generation. This approach, to the best of our knowledge, is pioneering in the field as it does not rely on pre-existing label combinations but instead creates new ones.

2. Within the framework of LCG, we propose a novel method named Tail-Driven Conditional Augmentation. TDCA enables the generation of balanced label combinations through the construction of a dual-weighted label graph and employs tail-driven label sampling based on the Metropolis-Hastings algorithm. Moreover, it utilizes contrastive label-conditioned generation to produce augmented texts that are both representative of the assigned labels and coherent with the original dataset.

3. Our experiments conducted across three datasets show that TDCA significantly reduced the long-tail effect in MLTC. The results show notable improvements in PSP@1, P@10, and N@10, with increases of **100.21%**, **16.58%**, and **11.65%**, respectively.

2 Tail-Driven Conditional Augmentation

Multi-label Text Classification (MLTC) aims to assign a subset of labels $Y \subseteq L$ to each instance x, where $L = \{l_1, l_2, ..., l_N\}$ represents the entire possible label space. This task can be formalized as learning a mapping function $f(x) \rightarrow 2^L$, predicting the power set of L. The approach of TDCA is straightforward: It commences with tail-driven label sampling to create label combinations and proceeds to contrastive label-conditioned generation, synthesizing text that aligns with these labels.

2.1 Tail-Driven Label Sampling

In MLTC, two primary characteristics emerge: (1) Each instance is associated with multiple labels, which exhibit significant correlations; (2) As depicted in Table 2, multi-label datasets encounter a pronounced long-tail effect, with tail labels dominating the label space yet being represented in only a small fraction of training instances. To address this, a balanced and correlation-aware label sampling approach is necessary. Consequently, we present Tail-Driven Label Sampling, which incorporates the Dual-Weighted Label Graph and Metropolis-Hastings sampling.

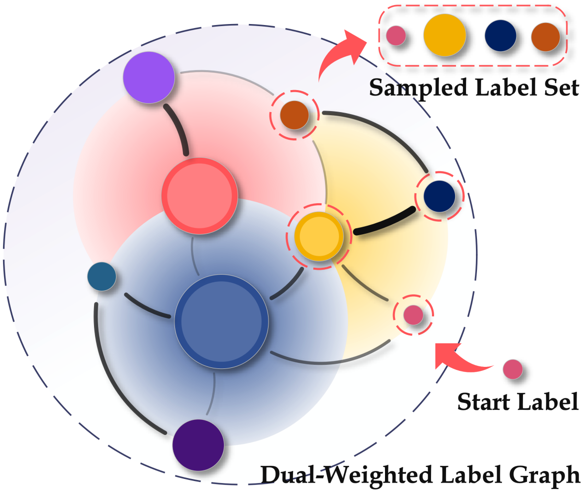


Figure 3: Tail-driven label sampling. Vertex size shows label frequency in the dataset, while edge thickness and color depth indicate label co-occurrence frequency.

2.1.1 Dual-Weighted Label Graph

The Dual-Weighted Label Graph (DWLG) is formally defined as $G = (V, E, W_v, W_e)$. It comprises vertices V, each signifying a distinct label, and edges E, linking pairs of vertices based on their co-occurrence in the dataset. Vertex weights are consolidated in $W_v = \{w_v(i) \mid i \in V\}$, with $w_v(i)$ representing the occurrence frequency of label i. The edge weights, denoted as $w_e(i, j)$ in W_e, capture the co-occurrence strength between labels i and j. This dual-weighted architecture yields an integrated perspective of both individual label properties and their interconnections, fundamental for advanced label sampling and data exploration[1].

2.1.2 Metropolis-Hastings Label Sampling

Employing the principles of Markov chain theory, Metropolis-Hastings (M-H) sampling (Metropolis et al., 1953; Hastings, 1970) can adeptly adjust to the target distribution, which is perfectly in tune with our aim to achieve a more balanced sampling strategy in long-tail distributions. We initiate M-H sampling of the DWLG by starting from a tail label, then iteratively transitions to other labels, directed by both the transition kernel and acceptance rate. This procedure is maintained until an adequate count of labels is sampled or a pre-determined step limit is reached.

Firstly, the transition kernel $q(i \rightarrow j)$ calculates the likelihood of moving from the current label i

[1] Our DWLG diverges from the traditional Label Correlation Graph concept introduced by Mittal et al. (2021), departing from random walk-based graph construction and incorporating dual weights for a clearer depiction of label dynamics.

Algorithm 1 Tail-Driven Label Sampling

```
Input:
    $G$: Dual-Weighted Label Graph
    $start$: Starting Label
    $steps$: Number of M-H Steps
    $T$: Temperature Parameter
    $maxLabels$: Max Labels to Collect
Output:
    $sampled$: Sampled Labels Set

$sampled \leftarrow \{start\}$
$current \leftarrow start$
for $i \leftarrow 1$ to $steps$ do
    $N \leftarrow$ GetNeighbors($current$, $G$)
    $next \leftarrow$ SampleNext($N$)
    $accept \leftarrow$ Acceptance($current$, $next$, $T$)
    if $accept$ then
        $current \leftarrow next$
        $sampled \leftarrow sampled \cup \{next\}$
    end if
    if $|sampled| \geq maxLabels$ then
        break
    end if
end for
return $sampled$
```

to an alternative label j within the DWLG during sampling. Considering the inter-label correlations, we define $q(i \to j)$in the following manner:

$$q(i \to j) \quad = \quad \frac{e^{w_e(i,j)}}{\sum_{k \in \text{neighbors}(i)} e^{w_e(i,k)}}, \tag{1}$$

where $w_e(i,j)$ denotes the edge weight between labels i and j in the DWLG.

Then, drawing upon the principles of information entropy (Shannon, 1948), the target distribution $p(i)$ is defined to encapsulate the significance and scarcity of label i:

$$p(i) = \frac{e^{s(i)/T}}{\sum_k e^{s(k)/T}}. \tag{2}$$

Here, $s(i) = -\log(w_v(i))$ is the importance score of label i, and T is a temperature parameter that moderates the distribution's smoothness.

Finally, the acceptance rate $\alpha(i \to j)$ evaluates whether to accept the transition from label i to label j, thereby steering the sampling outcomes towards the target distribution:

$$\alpha(i \to j) = \min\left(1, \frac{p(j) \cdot q(j \to i)}{p(i) \cdot q(i \to j)}\right). \tag{3}$$

2.2 Contrastive Label-conditioned Generation

Conditional generation, recognized for its capability to produce texts that are both diverse and controlled, has improved data augmentation in binary and multi-class scenarios. However, previous studies have largely concentrated on the label-conditioned duplication of existing data points. The true potential of conditional generation extends far beyond this practice.

Integrating LLMs into conditional generation introduces several challenges. When dealing with a sampled labels set: (1) An excess of labels leads to extended input sequences, complicating LLMs' ability to reflect each label in the generated text, often causing omissions; (2) LLMs' sensitivity to prompt selection results in erratic text generation quality; (3) Existing LLMs, enhanced with RLHF, tend to generate superfluous explanatory content, which can be counterproductive for data augmentation; (4) Texts generated by LLMs exhibit stylistic discrepancies with the original dataset.

We fine-tune LLM on the original dataset to mitigate these issues. Inspired by Song et al. (2023) and Rafailov et al. (2023), we devised two targeted loss functions for multi-label classification dataset augmentation: Label Match Loss ($\mathcal{L}_{\mathcal{LM}}$), ensuring generated texts align with each input label, and Style Consistency Loss ($\mathcal{L}_{\mathcal{SC}}$), which aids in producing texts that are coherent, controllable, and in stylistic harmony with the original dataset.

2.2.1 Style Consistency Loss

Initially, a subset $\{X^1, Y^1; \ldots; X^n, Y^n\}$ is randomly extracted from the training set, wherein $X = \{x_1, \ldots, x_{|X|}\}$ denotes a text, comprising a series of tokens x, and $Y = \{y_1, \ldots, y_{|Y|}\}$ represents a set of multiple labels associated with X. Regarding Y, we concat it with a prompt to create a composite text $c(Y)$, such as "Generate text for the labels $[y_1, \ldots, y_{|Y^i|}]$" to serve as an input for the LLM. To ensure that the text X_{aug} produced by the LLM aligns with X, the Style Consistency Loss is as follows:

$$\mathcal{L}_{\mathcal{SC}} = -\sum_t \log P_\phi(x_t | c(Y), x_{1,\cdots,t-1}). \tag{4}$$

Here, ϕ signifies the parameters of the LLM. Given the input $c(Y)$ and the prior tokens $x_{1,\cdots,t-1}$, our goal is to fine-tune the model with greater probability to predict the next token x_t.

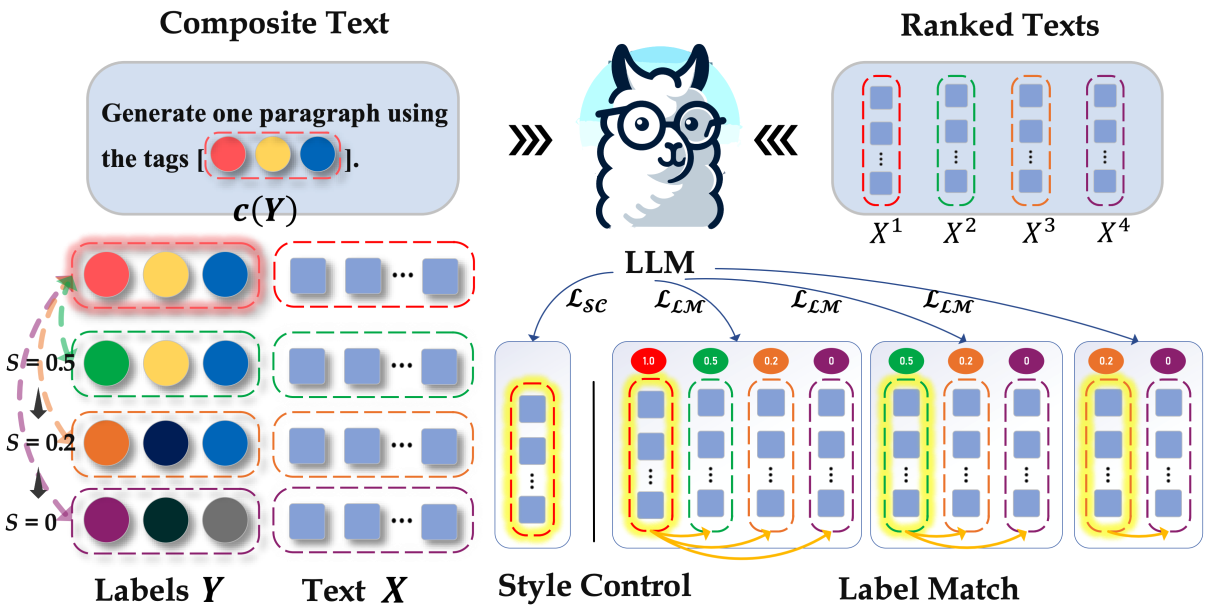

Composite Text
Generate one paragraph using the tags [ ].
$c(Y)$
»
LLM
«
Ranked Texts
X^1
X^2
X^3
X^4
$\mathcal{L}_{\mathcal{SC}}$
$\mathcal{L}_{\mathcal{LM}}$
$\mathcal{L}_{\mathcal{LM}}$
$\mathcal{L}_{\mathcal{LM}}$
S = 0.5
S = 0.2
S = 0
1.0
0.5
0.2
0
0.5
0.2
0
0.2
0
Labels Y
Text X
Style Control
Label Match

Figure 4: Fine-tuning LLM in Contrastive Label-conditioned Generation. $\mathcal{L}_{\mathcal{LM}}$ ensures alignment of generated texts with input labels. $\mathcal{L}_{\mathcal{SC}}$ supports the production of similar texts with the original dataset.

2.2.2 Label Match Loss

To effectively align the augmented input X_{aug} with the corresponding label combinations Y, we not only employ $\mathcal{L}_{\mathcal{SC}}$ to illustrate ideal generation but also aim to guide the LLM in distinguishing between good and bad ones. In the randomly selected subset of the training set $\{\boldsymbol{X^1}, \boldsymbol{Y^1}; X^2, Y^2; \ldots; X^n, Y^n\}$, where each Y^i is unique, the Jaccard similarity (Jaccard, 1912) is utilized to evaluate and rank the degrees of similarity, ranging from Y^1 and Y^2, to Y^1 and Y^n. For labels Y^1, its associated text X^1 is considered a **positive** example, while texts ranging from X^2 to X^n are deemed **negative** examples, exhibiting progressively higher degrees of dissimilarity.

We generalize the InfoNCE loss (van den Oord et al., 2018) using the Plackett-Luce model (Plackett, 1975; Luce, 1959), deriving the Label Match Loss as follows:

$$\mathcal{L}_{\mathcal{LM}} = -\sum_{i=1}^{n-1} \log \frac{\exp\left(\frac{r_\phi(X^i)}{\mathcal{T}_i^i}\right)}{\sum_{j=i}^{n} \exp\left(\frac{r_\phi(X^j)}{\mathcal{T}_i^j}\right)}, \qquad (5)$$

where ϕ denotes the parameters of the LLM, $\mathcal{T}$ represents the temperature coefficient, and $r_\phi(X^i)$ indicates the likelihood of the LLM generating X^i. For the set $\{\boldsymbol{X^1}, \boldsymbol{Y^1}; X^2, Y^2; \ldots; X^n, Y^n\}$, we initially compare X^1 with each of $X^2, \ldots, X^n$, followed by a comparison of X^2 with $X^3, \ldots, X^n$, aiming to align the LLM-generated X_{aug} with the label Y. Specifically, $r_\phi(X)$ is defined as:

$$r_\phi(X) = \frac{1}{|X|} \sum_{t=1}^{|X|} \log P_\phi(x_t | c(Y^1), x_{1,\cdots,t-1}), \qquad (6)$$

where X encompassing tokens $x_1, \cdots, x_{|X|}$, with $c(Y^1)$ representing the label of the positive sample integrated with the prompt.

Each comparison involves modulation of suppression for negative samples at varying degrees through temperature coefficients $\mathcal{T}$:

$$\mathcal{T}_i^{j>i} = \frac{1}{s(Y^1, Y^i) - s(Y^1, Y^j)}, \qquad (7)$$

$$\mathcal{T}_i^i = \min_{j>i} \mathcal{T}_i^j. \qquad (8)$$

Here, $s(Y^i, Y^j)$ denotes the Jaccard similarity between the label sets Y^i and Y^j.

2.2.3 Optimization Objective

Combining the label match loss and style consistency loss, the final loss function is:

$$\mathcal{L} = \mathcal{L}_{\mathcal{LM}} + \lambda \mathcal{L}_{\mathcal{SC}}. \qquad (9)$$

The hyperparameter λ balances the importance of label matching and style consistency. Optimizing $\mathcal{L}$ ensures that the generated text is not only

Dataset	N_{Train}	N_{Test}	W_{Avg}	L_{Avg}	L_{Total}	$L_{>100}$	$L_{<10}$
Eurlex-4K	15,449	3,865	1,237.88	5.32	3,956	233	2,396
MeSH-12K	9,996	3,500	178.47	12.27	12,784	211	9,951
Wiki10-30K	14,145	6,616	2,086.01	18.37	29,973	461	25,178

Table 2: Summary of datasets in Eurlex-4K, MeSH-12K, and Wiki10-30K. $\boldsymbol{N_{Train}}$: Number of training instances, $\boldsymbol{N_{Test}}$: Number of test instances, $\boldsymbol{W_{Avg}}$: Average words per instance, $\boldsymbol{L_{Avg}}$: Average labels per instance, $\boldsymbol{L_{Total}}$: Total number of labels, $\boldsymbol{L_{>100}}$: Labels with more than 100 instances, $\boldsymbol{L_{<10}}$: Labels with less than 10 instances.

relevant to the sampled labels Y but also stylistically coherent with the actual text X, striking a balance between accuracy and authenticity.

3 Experiments

3.1 Datasets

We employ three benchmark datasets: Eurlex-4K, MeSH-12K, and Wiki10-30K ("K" denotes the label count within each dataset). **Eurlex-4K** (Loza Mencía and Fürnkranz, 2008) comprises a corpus of European Union legal documents. We utilize its raw text, as provided in Ye et al. (2020), without applying stemming or stop-word removal. **MeSH-12K**, a subset culled from the BioASQ datasets[2](Tsatsaronis et al., 2015), is composed of article titles and abstracts from PubMed, annotated with Medical Subject Headings (MeSH) as their labels. **Wiki10-30K**, originating from Wikipedia, is a refined iteration of Wiki10-31K[3] (Zubiaga, 2012). While prior studies merely utilized the label IDs in Wiki10-31K for classification, we undertook a meticulous review and cleansing of this dataset. This entailed the exclusion of content-lacking instances and labels constituted solely of punctuations (*e.g.*, ".", "!", "!!!") or NLTK-listed stopwords (*e.g.*, "and", "or"). More statistical information can be found in Table 2.

3.2 Evaluation Metrics

To assess the performance of each method, we use the following evaluation metrics: $P@k$, $PSP@k$, and $N@k$.

Precision at k ($P@k$) measures the precision of the top k predicted labels in matching the true labels. For an instance with its top-k predicted label set $\hat{Y}@k$ and actual label set Y, $P@k$ is defined as:

$$P@k = \frac{|\hat{Y}@k \cap Y|}{k}, \qquad (10)$$

where $|\hat{Y}_k \cap Y|$ is the number of correct predictions in the top k labels. On the Wiki10-30K, if a model accurately predicts the head labels ($L_{>100} = 461$) and ignores the rest, it would achieve a $P@1$ of 90.8%. This high precision, however, overlooks the informative tail labels.

Propensity Scored Precision at k ($PSP@k$) mitigates this limitation by modifying precision to give higher weights to less frequent labels. Given the top-k predicted label set $\hat{Y}_k$ and the actual label set Y, $PSP@k$ is defined as:

$$PSP@k = \frac{\sum_{y \in \hat{Y}@k \cap Y} \frac{1}{\pi_y}}{k}. \qquad (11)$$

Here, π_y represents the propensity score of label y, quantifying the likelihood of encountering label y in the training dataset.

Normalized Discounted Cumulative Gain at k ($N@k$) evaluates the ranking quality of predicted labels, considering both the relevance of the labels and their ranking positions. The Discounted Cumulative Gain (DCG) for a predicted label $\hat{y}_i$ at rank i is calculated as:

$$DCG@k = \sum_{i=1}^{k} \frac{\mathbb{I}(\hat{y}_i \in Y)}{\log_2(i+1)}, \qquad (12)$$

where $\mathbb{I}(\cdot)$ is an indicator function. $N@k$ is the ratio of $DCG@k$ to $iDCG@k$:

$$iDCG@k = \sum_{i=1}^{\min(k,|Y|)} \frac{1}{\log_2(i+1)}. \qquad (13)$$

$$N@k = \frac{DCG@k}{iDCG@k}. \qquad (14)$$

3.3 Experiment Settings

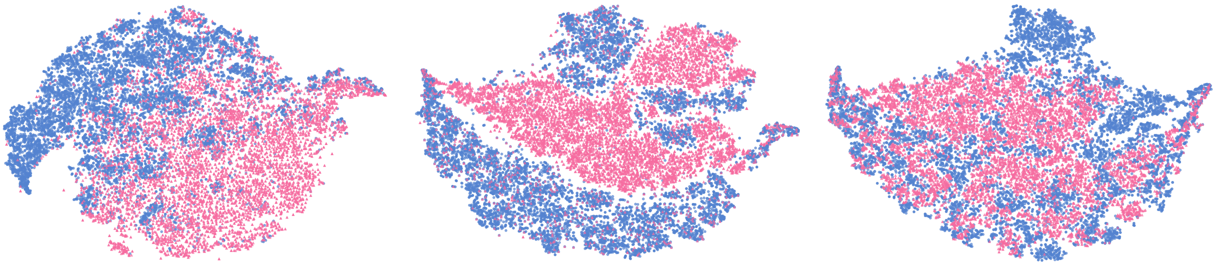
We compare TDCA with three data augmentation baselines: EDA, BT, and CA. EDA (Wei and Zou, 2019) involves synonym replacement and random insertion/swap/deletion. In BT, we

[2] http://participants-area.bioasq.org

[3] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html

Dataset	Methods	E/R	PSP@1	PSP@3	PSP@5	P@1	P@3	P@5	P@10	N@3	N@5	N@10
EurLex	*Raw*	1	*42.65*	*50.46*	*53.89*	*85.56*	*73.11*	*60.91*	*38.35*	*76.48*	*70.46*	*73.82*
	EDA	1+100%	42.33	50.50	54.53	85.54	72.77	60.84	38.33	76.18	70.33	73.78
	BT	1+25%	42.77	50.73	54.42	86.03	73.42	61.15	38.61	76.84	70.79	74.34
	CA	1+150%	43.53	52.30	55.89	86.70	74.04	61.66	38.93	77.48	71.33	74.86
	TDCA	**1+400%**	**49.67**	**57.87**	**61.19**	**88.15**	**76.49**	**64.04**	**40.68**	**79.75**	**73.79**	**77.70**
MeSH	*Raw*	1	*18.04*	*21.56*	*23.32*	***90.46***	*74.64*	*63.61*	*46.04*	*78.32*	*69.84*	*58.44*
	EDA	1+200%	18.13	21.93	24.04	89.31	74.65	63.66	45.90	78.06	69.70	58.20
	BT	1+200%	18.16	21.66	23.64	89.20	73.89	62.70	45.18	77.45	68.91	57.49
	CA	1+500%	19.33	24.14	26.65	90.03	77.85	66.99	49.07	80.74	72.63	61.34
	TDCA	**1+1000%**	**35.84**	**42.08**	**46.45**	90.20	**82.67**	**75.61**	**59.92**	**84.42**	**79.23**	**70.57**
Wiki	*Raw*	1	*16.22*	*16.66*	*17.33*	***87.89***	*77.42*	*68.06*	*51.79*	*79.84*	*72.77*	*60.47*
	EDA	1+50%	16.67	17.30	18.13	87.24	76.43	67.00	50.82	78.92	71.78	59.52
	BT	1+50%	16.66	17.50	18.28	87.32	76.98	67.59	51.42	79.37	72.29	60.07
	CA	1+200%	17.83	19.15	20.34	87.09	76.48	67.52	51.20	78.97	72.16	59.89
	TDCA	**1+3000%**	**46.31**	**39.02**	**36.92**	87.12	**78.27**	**72.03**	**58.79**	**80.31**	**75.48**	**65.88**

Table 3: Performance comparison of DA techniques (EDA, BT, CA, TDCA) against raw dataset on Wiki10-30K, EurLex-4K, and MeSH-12K. Metrics for each method are reported at their **optimal** E/R (Expansion Ratio).



(a) Back-Translation

(b) EDA

(c) TDCA

Figure 5: A t-SNE (van der Maaten and Hinton, 2008) visualization of the MeSH-12K dataset comparing various text data augmentation methods: (a) BT, (b) EDA, and (c) our proposed TDCA, where ● represents original data, and ▲ represents augmented data.

select French, Chinese, Russian, Italian, and Spanish as intermediate languages for English paraphrasing via `nllb-200-distilled` model (Costa-jussà et al., 2022). CA and TDCA utilize LLaMa-based LLM, `Qwen-7B-Chat` (Bai et al., 2023), for label-conditioned text generation. For MLTC, we utilized LightXML (Jiang et al., 2021) with `bert-base-uncased` (Devlin et al., 2019). The EurLex-4K setup included a 1e-4 learning rate, 15 epochs, batch size of 16, and max token length of 512, with SWA (Izmailov et al., 2018) applied post 10 epochs (step size 200). For Wiki10-30K, we extended training to 30 epochs (SWA step size 300), maintaining other parameters. MeSH-12K followed the Wiki10-30K settings, with a batch size of 8 and a max token length of 256. TDCA utilized Metropolis-Hastings sampling from labels with <100 instances, with a 1000-step limit and temperature of 10. Fine-tuning involved a λ value of 0.2 for loss balance, 2 epochs, a 512 sequence length, and a 5e-6 learning rate. All experiments were conducted on 8 NVIDIA A100 GPUs.

4 Results

4.1 Performance Comparison

Performance assessments were carried out on the EurLex-4K, MeSH-12K, and Wiki10-30K on various expansion ratios to evaluate the effectiveness of our proposed TDCA compared to EDA, BT, and CA. As Table 3 indicates, all methods mitigated the long-tail effect in MLTC datasets to varying degrees: EDA and BT marginally enhanced $PSP@k(k = 1, 3, 5)$ by 1.89% and 1.95%, respectively. CA's diverse text generation led to a 9.45% increase, limited by unchanged labels. TDCA, integrating tail-driven sampling, impressively reduced the long-tail effect with an 85.61% improvement.

Moreover, data augmentation's effectiveness correlates with the number of dataset labels. TDCA enhanced $PSP@1$ by 16.46% on EurLex-4K, 98.78% on MeSH-12K, and 185.51% on Wiki10-30K. This result aligns with the expectation that more labels intensify the long-tail effect, enhancing the utility of data augmentation.

		EDA			BT			CA w/o F-T			CA (TDCA w/o M-H)			TDCA w/o F-T			TDCA		
		PSP	P	N	PSP	P	N	PSP	P	N	PSP	P	N	PSP	P	N	PSP	P	N
EurLex	1+25%	53.94	60.58	70.04	54.42	61.15	70.79	54.85	61.53	71.08	54.42	61.17	70.74	54.06	61.05	70.77	55.61	61.73	71.31
	1+50%	54.25	60.94	70.46	53.57	60.26	69.95	55.21	61.61	71.18	54.51	61.07	70.71	54.28	61.16	71.03	56.57	62.03	71.77
	1+100%	54.53	60.84	70.33	53.97	60.63	70.12	55.15	61.31	70.81	55.39	61.50	71.21	55.73	61.92	71.82	58.97	63.37	73.17
	1+200%	53.60	59.91	69.14	53.85	60.42	69.82	56.04	61.76	71.39	56.36	62.01	71.55	57.21	62.76	72.59	60.31	63.72	73.36
	1+500%	54.47	60.40	69.46	53.96	60.51	69.90	56.42	61.92	71.52	56.43	62.00	71.50	60.77	63.75	73.45	61.21	64.08	73.63
MeSH	1+25%	23.34	63.12	69.37	23.55	63.66	69.67	23.95	63.94	70.02	24.27	64.25	70.36	24.38	64.20	70.15	24.70	64.65	70.55
	1+50%	23.64	63.31	69.44	23.67	63.41	69.61	24.64	64.65	70.63	24.68	65.07	70.95	24.98	64.78	70.69	25.80	65.65	71.44
	1+100%	23.90	63.27	69.38	23.60	63.07	69.28	25.30	65.09	71.04	25.31	65.61	71.34	26.65	66.15	71.78	27.61	66.93	72.28
	1+200%	24.04	63.66	69.70	23.64	62.70	68.91	25.96	65.69	71.52	26.16	66.52	72.14	29.24	68.19	73.49	31.60	69.86	74.71
	1+500%	24.50	64.31	69.88	23.30	61.81	67.93	26.36	66.09	71.70	26.65	66.99	72.63	36.22	72.21	76.58	38.79	73.76	78.00
Wiki	1+25%	17.89	67.73	72.42	17.95	68.03	72.61	18.16	68.33	72.93	18.14	68.40	73.01	17.73	68.01	72.75	17.85	68.54	73.17
	1+50%	18.13	67.00	71.78	18.28	67.59	72.29	18.73	68.24	72.80	18.80	68.27	72.82	18.04	68.12	72.77	18.39	68.69	73.31
	1+100%	18.23	65.92	70.58	18.35	66.43	71.26	19.50	67.74	72.34	19.63	67.82	72.34	19.03	68.86	73.46	19.17	69.07	73.66
	1+200%	18.17	65.00	69.68	18.03	64.74	69.54	20.03	66.89	71.57	20.34	67.52	72.16	20.97	69.51	73.80	21.38	70.24	74.49
	1+500%	17.83	63.99	68.64	17.51	63.92	68.53	19.85	65.07	69.73	20.64	66.84	71.39	25.75	70.79	74.73	26.35	70.98	74.91

Table 4: Performance comparison of EDA, BT, TDCA, and three TDCA variants: CA (TDCA w/o M-H), TDCA w/o F-T, and CA w/o F-T across various expansion ratios in terms of $PSP@5$, $P@5$, and $N@5$. In the table, deeper shades of red indicate higher values, while deeper shades of green denote lower values.

Regarding $P@K$ and $N@K (k = 1, 3, 5, 10)$, label-conditioned augmentation methods, CA and TDCA, consistently improved performance, with CA increasing $P@K$ by 1.42% and $N@K$ by 1.09% on average, and TDCA by 8.16% and 5.95%. EDA and BT, however, had a marginal negative impact: EDA decreased $P@k$ and $N@K$ by 0.63% and 0.64%, while BT showed similar declines by 0.52% and 0.49%. This is further supported by t-SNE visualizations (Figure 5), where TDCA's augmented data demonstrate better integration with original data, unlike EDA and BT.

TDCA not only effectively counters the long-tail effect in MLTC datasets but also bolsters prediction precision and ranking quality. The improvements are progressive with larger values of k in all metrics, which indicates that label-conditioned augmentation is more adaptable to MLTC's extremely large label space. Furthermore, the quantitative experiments (Table 4) reveal that TDCA outperforms traditional methods at equivalent ratios. While EDA and BT exhibit increasingly negative effects with more augmented data, TDCA's benefits progressively amplify. For comprehensive details, refer to training logs in Appendix B.

The case study in Appendix A uncovers an intriguing aspect: the original dataset often exhibits an implicit link between labels and text, significantly challenging the learning process of models. The LLM-based CA and TDCA render these labels more explicit in the generated texts, facilitating an easier learning of correspondences.

4.2 Ablation Analysis

Table 4 presents the results of TDCA and its three variant models under different expansion ratios. These variants are: CA (TDCA without M-H based tail-driven sampling, *i.e.*, TDCA w/o M-H), TDCA w/o F-T (TDCA without contrastive label-conditioned fine-tuning), and CA w/o F-T (TDCA w/o M-H & F-T).

The varying shades of red (high) and green (low) in the table illustrate that TDCA and TDCA w/o F-T significantly outperform both CA (TDCA w/o M-H) and CA w/o F-T (TDCA w/o F-T & M-H) across all metrics. This highlights the pivotal role of M-H based tail-driven sampling in TDCA, validating our proposed Label Creative Generation. Furthermore, the comparison between TDCA and TDCA w/o F-T, as well as between CA and CA w/o F-T, confirms the effectiveness of contrastive fine-tuning. Overall, these results demonstrate that each component of the TDCA structure is effective and contributes to its overall performance.

5 Conclusion

In this paper, we introduce a new multi-label data augmentation paradigm named Label Creative Generation (LCG). Under LCG, we propose Tail-Driven Conditional Augmentation (TDCA). TDCA facilitates the creation of balanced label combinations by dual-weighted label graph and tail-driven label sampling. Furthermore, TDCA employs a contrastive label-conditioned generation to produce augmented texts that match each label and maintain consistency with the original dataset. Empirical evaluations on three datasets with varying label counts demonstrate the effectiveness of TDCA. Our approach significantly surpasses existing data augmentation methods, effectively mitigates the long-tail effect, and enhances model prediction performance in MLTC.

Limitations

- This method is primarily effective for textual labels with semantic content. It is less suited for labels represented by numerical IDs or non-descriptive identifiers.
- Compared to EDA and other non-model-based data augmentation techniques, our approach is more resource-intensive. However, its resource consumption is comparable to seq2seq model-based methods like back-translation.
- Due to computational resource constraints, we were unable to fully explore the upper limits of TDCA's performance improvements. On the EurLex dataset, we observed optimal expansion ratios between 400% to 500%. Beyond this range, performance tended to plateau or slightly decrease. In contrast, on the MeSH-12K and Wiki10-30K datasets, with more extensive labels, we experimented with expansion ratios up to 1000% and 3000%, respectively, without reaching an apparent performance ceiling.
- Our exploration did not extend to closed-source LLMs such as ChatGPT, Bard, or Claude, limited by API access. Nonetheless, considering the promising results achieved with the un-tuned Qwen-7B model (*i.e.*, TDCA w/o F-T in ablation study), we believe that employing closed-source LLMs via API calls can yield comparable or superior results.
- We observed limitations in fine-tuning (FT), potentially due to the simplicity of our data synthesis task (generating text from labels), where using prompts alone could teach a LLM this task. However, we believe that the LCG extends beyond this. For instance, in recommendation systems, LCG could be used on the MovieLens dataset to generate diverse user information for movie recommendations, such as address, age, postal code, etc. Nevertheless, relying solely on prompts for an LLM might fall short, whereas FT ensures this capability.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62206192); the Natural Science Foundation of Sichuan Province (No. 2023NS-36 FSC1408); the Science and Technology Major Project of Sichuan Province; and the Fundamental Research Funds for the Central Universities (No. 1082204112364). We sincerely thank the ARR 2023 Dec reviewers for their insightful feedback and extend our great gratitude to Yang Li and Guanyu Qiao for the stimulating discussions at the inception of this research, which significantly enriched this work.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*, abs/2309.16609.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2023. A survey on data augmentation for text classification. *ACM Comput. Surv.*, 55(7):146:1–146:39.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *CoRR*, abs/2303.12712.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti

Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *CoRR*, abs/2207.04672.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*, v3.

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.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 489–500. Association for Computational Linguistics.

Zhiqi Ge, Yuanyuan Guan, Ximing Li, and Bo Fu. 2022. Consistent, balanced, and overlapping label trees for extreme multi-label learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17-21, 2022*, pages 551–560. ACM.

W. K. Hastings. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 876–885. AUAI Press.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50.

Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 935–944. ACM.

Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. 2021. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 7987–7994. AAAI Press.

Jan Kocon, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocon, Bartlomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Milkowski, Marcin Oleksy, Maciej Piasecki, Lukasz Radlinski, Konrad Wojtasik, Stanislaw Wozniak, and Przemyslaw Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Inf. Fusion*, 99:101861.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6449–6464. Association for Computational Linguistics.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7056–7066, Online. Association for Computational Linguistics.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 115–124. ACM.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9031–9041. Association for Computational Linguistics.

Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023. Breaking the bank with ChatGPT: Few-shot text classification for finance. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 74–80, Macao. Association for Computational Linguistics.

Eneldo Loza Mencía and Johannes Fürnkranz. 2008. Efficient pairwise multilabel classification for large-scale problems in the legal domain. In *Machine*

Learning and Knowledge Discovery in Databases, pages 50–65, Berlin, Heidelberg. Springer Berlin Heidelberg.

R. D. Luce. 1959. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Anshul Mittal, Noveen Sachdeva, Sheshansh Agrawal, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. ECLARE: extreme classification with label graph correlations. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 3721–3732. ACM / IW3C2.

R. L. Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *CoRR*, abs/2305.18290.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2023. Preference ranking optimization for human alignment. *CoRR*, abs/2306.17492.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artieres, Axel Ngonga, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jens Van Nooten and Walter Daelemans. 2023. Improving Dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270, Toronto, Canada. Association for Computational Linguistics.

Jason W. Wei and Kai Zou. 2019. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 6381–6387. Association for Computational Linguistics.

Jie Xiong, Li Yu, Xi Niu, and Youfang Leng. 2023. Xrr: Extreme multi-label text classification with candidate retrieving and deep ranking. *Information Sciences*, 622:115–132.

Hui Ye, Zhiyu Chen, Da-Han Wang, and Brian D. Davison. 2020. Pretrained generalized autoregressive model with adaptive probabilistic label clusters for extreme multi-label text classification. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 10809–10819. PMLR.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5812–5822.

Arkaitz Zubiaga. 2012. Enhancing navigation on wikipedia with social tags. *CoRR*, abs/1202.5469.

A Case Study

MeSH terms (labels)

Community Health Planning; Consultants; **Data Collection**; Delivery of Health Care, Integrated; **Evaluation Studies as Topic**; Health Care Surveys; **Health Plan Implementation**; Health Services Needs and Demand; Iowa; Public Health Administration; State Health Plans; **United States**

ORIGINAL

Determinants of successful community health system planning. Needs assessment and other community health planning projects are being conducted in many parts of the country. Whether these projects are successful, in that they lead to implementation of recommendations, has not been systematically assessed. This study describes how local public health officials in Iowa view the planning projects that were conducted in their communities. Planning projects occurred in over half of the counties in the state. Most were led by local public health agencies. Hospitals were participants in three-fourths of the projects. Successful implementation was associated with the number of planning techniques used and the breadth of representation on the steering committee. The most common recommendations related to the expansion of prevention/promotion services and community health nursing. Public health nursing had been integrated with local hospitals in one-third of the counties, with no reductions in performance.

EDA

determinants of the community health system other needs assessment and planning community health planning projects are being conducted in many parts of the country whether health the are successful assessed that they lead to implementation of recommendations has not been on in related study describes how local public health officials in iowa view agencies planning projects that were conducted in their systematically planning projects services in over half of the counties in the state most successful led by local public techniques this hospitals were participants in three fourths of the projects successful implementation was associated with third number of planning these used and the breadth of representation communities the steering committee projects most common recommendations the to the expansion of prevention promotion occurred and community health nursing public health nursing had been integrated with local hospitals in one the of were counties with no reductions in performance

BT

Determinants of successful community health system planning. Needs assessment and other community health planning projects are being carried out in many parts of the country. The success of these projects, because they lead to the implementation of recommendations, has not been systematically evaluated. This study describes how local public health officials in Iowa view planning projects in their communities. More than half of the counties in the state have planned projects. Most of them are led by local public health institutions. Hospitals are involved in three-quarters of the projects. The successful implementation is related to the number of planning technologies used and the representativeness of the steering committee. The most common suggestions involve expanding prevention/publicity services and community health care. In one third of counties, public health care has been integrated with local hospitals, but the performance has not declined.

CA

Evaluation of Community Health Planning Consultants in Iowa: A Survey of Data Collection and Delivery of Integrated Health Care Services. Community health planning consultants play a critical role in the implementation of state health plans and addressing the health services needs of communities. This study aimed to evaluate the effectiveness of community health planning consultants in Iowa in facilitating the delivery of integrated health care services. A survey was conducted among community health planning consultants to collect data on their roles, responsibilities, and experiences in implementing state health plans. The survey also assessed the challenges faced by consultants in collecting and analyzing data, delivering integrated health care services, and evaluating the impact of their interventions. The results showed that community health planning consultants in Iowa are effective in facilitating the delivery of integrated health care services, but face challenges in data collection and evaluation. The study recommends the development of standardized data collection tools and training programs for consultants to enhance their skills in data collection and analysis. Additionally, the study highlights the need for ongoing evaluation of the effectiveness of community health planning consultants in addressing the health services needs of communities and implementing state health plans.

Figure 6: Case study on the MeSH-12K dataset evaluating three traditional data augmentation methods: EDA, BT, and CA. Head labels (over 100 instances in training) are marked in red; tail labels (under 10 instances) in blue. The study finds that paraphrasing methods, EDA and BT, although effective in diversifying data, inadvertently add noise. CA-generated text is smooth and aligns with all MeSH terms (labels). Notably, the original dataset's text-label correspondence is often implicit, posing challenges in classifier training. For instance, **Consultants** are indirectly implicated through "local public health agencies" and "hospitals," and **Data Collection** is inferred from the context of health planning. The term **United States** is indirectly indicated by mentioning "Iowa." These nuances complicate classification, which is not simplified by mere paraphrasing. In contrast, CA's label-conditioned generation directly incorporates labels into the text, significantly aiding the classification of rare tail labels.

MeSH terms (labels)

Fanconi Anemia; Stress, Mechanical; Surface Properties; Syndrome; **Chromosomes, Human, Pair 13**; **Founder Effect**; **Mites**; **Thromboxane-A Synthase**; **Benzoquinones**; **Female**; **Thromboxane B2**; **Leukotrienes**

TDCA

Exploring the Role of Mechanical Stress and Surface Properties in Fanconi Anemia Syndrome: Insights from the Founder Effect, Chromosomes, and Mites. Fanconi Anemia (FA) is a rare genetic disorder characterized by chromosomal instability, bone marrow failure, and an increased risk of cancer. Previous studies have highlighted the involvement of oxidative stress in FA pathogenesis, but the role of mechanical stress and surface properties remains unclear. This review aims to examine the impact of mechanical stress and surface properties on FA, particularly in the context of the founder effect, chromosomes, and mites. We discuss the potential mechanisms through which these factors may contribute to FA, such as the activation of thromboxane-A synthase and the production of benzoquinones and leukotrienes. Additionally, we explore the role of gender in FA and the potential differential effects of these factors in female patients. By synthesizing current knowledge with insights from various research fields, this review sheds light on the complex interplay between mechanical stress, surface properties, and FA pathogenesis, providing a foundation for future investigations in this area.

MeSH terms (labels)

Electronics; **Blood Urea Nitrogen**; Income; Oregon; Personal Autonomy; Ethnic Groups; **Hypercholesterolemia**; Hospital Administration; Computers; Health Knowledge, Attitudes, Practice; **Research Subjects**; **Cardiac Care Facilities**

TDCA

The Association between Income and Health Outcomes among Various Ethnic Groups in Oregon: A Study of Personal Autonomy, Hypercholesterolemia, and Hospital Administration Using Electronics and Computers: This study aimed to investigate the relationship between income and various health outcomes among different ethnic groups residing in Oregon. The study utilized electronic data and computer technology to analyze data from multiple sources, including hospital administration records, research subjects' health knowledge, attitudes and practices, and cardiac care facilities. Specifically, the study examined the impact of income on personal autonomy, blood urea nitrogen levels, and the prevalence of hypercholesterolemia within different ethnic groups. Results showed a significant association between income and personal autonomy, with lower income groups experiencing higher levels of restriction in decision-making about their health. Additionally, income was found to be positively correlated with blood urea nitrogen levels and the prevalence of hypercholesterolemia, suggesting that higher income may play a protective role against these health concerns. The implications of these findings for hospital administrations and future research on health disparities among various ethnic groups are discussed.

MeSH terms (labels)

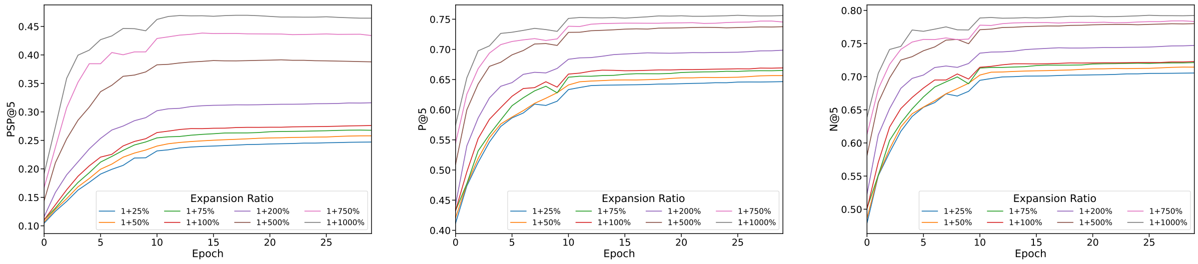
Polyethylene Glycols; **Bacterial Proteins**; Point Mutation; **Regulon**; Antigens, CD4; **Transplantation Immunology**; **Calorimetry**; Deoxycholic Acid; **T-Lymphocytes**; **Receptors, Lymphocyte Homing**; Surface Properties; Mice, Inbred CBA

TDCA

Investigating the Impact of Polyethylene Glycols on Bacterial Proteins and Point Mutation in the Regulon of CD4 Antigens in Transplantation Immunology using Calorimetry and Deoxycholic Acid in CBA Mice. This study aims to explore the effects of polyethylene glycols (PEGs) on bacterial proteins and point mutation in the regulon of CD4 antigens in transplantation immunology. Through the use of calorimetry and deoxycholic acid, we will investigate how PEGs alter the surface properties of T-lymphocytes and receptors involved in lymphocyte homing. The study will be conducted in CBA mice, a commonly used model for transplantation research. Results from this study will provide valuable insights into the role of PEGs in modifying the immune response, potentially leading to the development of targeted therapies for transplantation immunology.

Figure 7: Illustration of TDCA performance on the MeSH-12K dataset. Similar to Figure 6, red indicates head labels (with over 100 instances in the training set), and blue denotes tail labels (with fewer than 10 instances). But TDCA exhibits fewer red head labels and a greater number of blue tail labels. Despite the low frequency of occurrence, the labels are still related due to the construction of the Dual-Weighted Label Graph and Tail-Driven Label Sampling, as seen in the first example: **Fanconi Anemia** is linked to **Chromosomes, Human, Pair 13** due to mutations causing this inherited disease. Similarly, **Syndrome** is a term that includes conditions like **Fanconi Anemia**. The enzyme **Thromboxane-A Synthase** is crucial in producing **Thromboxane B2**. Furthermore, **Leukotrienes**, known for their role in inflammation, can interact with **Thromboxane B2** during certain physiological or pathological conditions. These interrelations facilitate the LLMs' task of generating coherent texts, and the augmented data is meaningful in real-world contexts.

B Training Logs

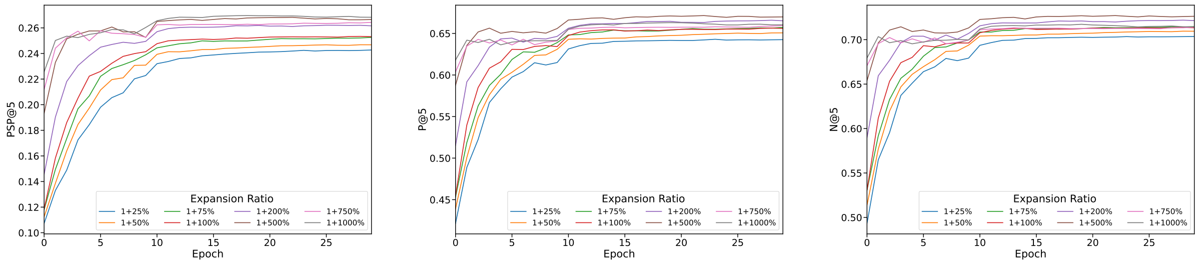


(a) PSP@5

(b) P@5

(c) N@5

Figure 8: Training logs of **TDCA** at different expansion ratios for PSP, P, and N metrics on MeSH-12K.

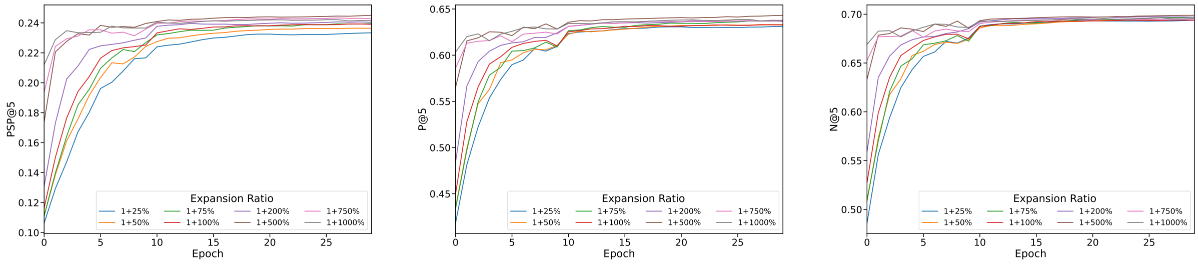


(a) PSP@5

(b) P@5

(c) N@5

Figure 9: Training logs of **CA** at different expansion ratios for PSP, P, and N metrics on MeSH-12K.

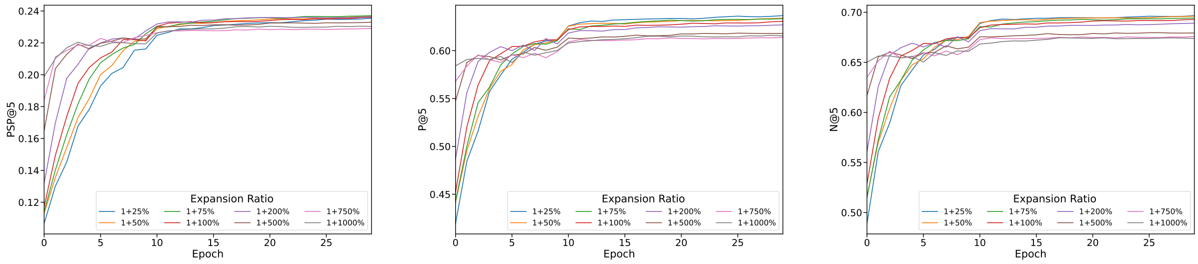


(a) PSP@5

(b) P@5

(c) N@5

Figure 10: Training logs of **EDA** at different expansion ratios for PSP, P, and N metrics on MeSH-12K.



(a) PSP@5

(b) P@5

(c) N@5

Figure 11: Training logs of **BT** at different expansion ratios for PSP, P, and N metrics on MeSH-12K.