Generating Efficient Training Data via LLM-based Attribute Manipulation

Anonymous ACL submission

Abstract

In this paper, we propose a novel method, Chain-of-Thoughts Attribute Manipulation (CoTAM), to guide few-shot learning by carefully crafted data from Large Language Models (LLMs). The main idea is to create data with changes only in the attribute targeted by the task. Inspired by facial attribute manipulation, our approach generates label-switched data by leveraging LLMs to manipulate task-specific attributes and reconstruct new sentences in a controlled manner. Instead of conventional latent representation controlling, we implement chain-of-thoughts decomposition and reconstruction to adapt the procedure to LLMs. Extensive results on text classification and other tasks verify the advantage of CoTAM over other LLM-based text generation methods with the same number of training examples. Analysis visualizes the attribute manipulation effectiveness of CoTAM and presents the potential of LLM-guided learning with even less supervision.[1]

1 Introduction

Recent large language models (LLMs) have demonstrated unprecedented few-shot learning capability when presented with in-context training examples (Brown et al., 2020; Wei et al., 2022c; Dong et al., 2022). Yet, such a capability is often combined with a large-scale model that is expensive for deployment (Brown et al., 2020; Chowdhery et al., 2022; OpenAI, 2023). Moreover, during inference time, the context containing demonstrations needs to be concatenated with every test input, which further increases computational burdens. Therefore, developing a small language model with the help of LLMs is one of the alternative solutions.

To effectively leverage the power of LLMs, previous research first generates new data with LLMs hinted by few-shot demonstrations, and then fine-tune a small pre-trained language model with the augmented dataset for better few-shot performance (Yoo et al., 2021; Sahu et al., 2022b; Dai et al., 2023; Lin et al., 2023). The small model can then be deployed offline without further LLM queries, and therefore improve inference efficiency. However, these methods usually prompt LLMs to wildly generate new examples without proper control, which hinders the informativeness of generated data and might induce spurious correlation. As shown Figure 1 (left), generated data without control have high variation and thus hard to be learned by small models.

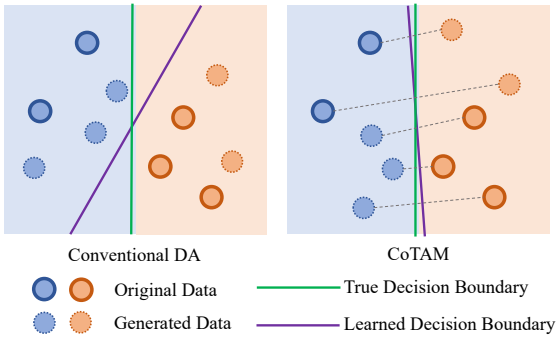

Figure 1: An illustrative comparison in case of binary classification. Conventional data augmentation generates uncontrolled data, while CoTAM directly reflects decision boundaries through task instructions. We present a real example in Figure 3.

In this paper, we investigate a more controlled and efficient generation. Our motivation is attribute manipulation in computer vision (Shen et al., 2020; Shen and Zhou, 2021) where attributes are manipulated in the latent space of the encoder to reconstruct new instances. A similar idea has been proposed in the language domain to manipulate task-specific semantics (e.g. sentiment) while preserving the original meaning of sentence (Kruengkrai, 2019a; Zhou et al., 2022). As shown in Figure 1, controlled attribute manipulation can efficiently find accurate decision boundaries since (1) we directly manipulate along the direction of

[1]Code: github.com/KomeijiForce/CoTAM

task-specific attributes. and (2) maintain the rest of the attributes as before.

Applying attribute manipulation for language is challenging due to two main reasons: (1) It is difficult to select a set of attributes. Sentences often contain various attributes (e.g. topic, emotion, intent) which might be different for each domain or dataset. Simply using a set of pre-defined attributes is labor-intensive and limited. (2) Reconstructing sentences with manipulated attributes requires a high-level understanding of semantics. Previous methods rely on random masking to reconstruct sentences (Zhou et al., 2022) which significantly degrades the diversity and plausibility of generated sentences.

To address the above challenges, we propose a chain-of-thoughts (Wei et al., 2022d) (CoT)-based method, Chain-of-Thoughts Attribute Manipulation (CoTAM), which utilizes an instruction-tuned LLM to manipulate attributes and reconstruct new sentences. Specifically, we prompt the LLMs following a three-step framework. In Step 1, we directly query LLMs to decompose sentences into multiple attributes that are independent of task-specific attributes. We argue that such a set of dynamic attributes capture the uniqueness of single sentences and fit for all domains without fine-tuning models. In Step 2, we instruct LLMs to output a guideline to switch task-specific attributes and maintain the others. Finally, in Step 3, we prompt the LLMs to reconstruct the sentence based on the guideline from Step 2. All these steps are conducted in a single query of LLM which guarantees the consistency of attribute manipulation and reconstruction. Furthermore, using LLMs benefits the interpretability of our framework where attributes are completely transparent to users.

We run CoTAM under few-shot settings in 4 natural language tasks including text classification, natural language inference, textual similarity and multiple choice question answering. We compare with strong baselines that utilize the same LLMs and generate the same amount of data. We evaluate quality of generated data via fine-tuning small language models. We further extend our evaluation to parameter-efficient methods such as nearest neighbor classifiers on text classification. Both results show significant and consistent performance improvements. Our analysis further demonstrates the importance of our design in CoTAM. The ablation study shows the importance of each thought in the CoT. The principal component analysis specifically depicts the effectiveness of LLMs in manipulating textual attributes to satisfy the task target. We further explore how to exploit CoTAM under even less supervision like fewer-shot, single-label, and LLM-proposed datasets. Results on the out-of-domain dataset show the generality of data generation from CoTAM. Finally, our case study presents how CoTAM manipulates real datasets.

Our contributions are three-fold: (i) We propose a novel LLM-guided few-shot learning procedure, CoTAM, which efficiently trains smaller models on data manipulated by LLMs. (ii) We conduct experiments on text classification and other tasks. Fine-tuning and instance-based results verify the advantage of our CoTAM. (iii) Our analysis specifies the advantages of CoTAM on different setups like less supervision and out-of-domain datasets.

2 Related Work

Attribute Manipulation aims to control certain attributes of the data. A general application of attribute manipulation is to change the visual attributes in facial images (Shen et al., 2020; Shen and Zhou, 2021). Image manipulation generally involves the transformation of image representations (Perarnau et al., 2016; Xiao et al., 2018; Shen et al., 2020) in the latent space. In natural language processing, the closest topic to attribute manipulation is data flipping (Kruengkrai, 2019b; Zhou et al., 2022), which replaces key spans in the text to switch its label. Obviously, many textual attributes like topics cannot be manipulated by span replacement. Thus, we choose to adapt the LLM to manipulate a latent space approximated by a series of attributes proposed by the LLM.

Controllable Generation is another close topic to our CoTAM. These methods typically generate texts from a continuous latent space discretely by controlling certain dimensions (Bowman et al., 2016; Hu et al., 2017; Yang and Klein, 2021). The controllable generator is trained by maximizing a variational lower bound on the data log-likelihood under the generative model with a KL divergence loss (Hu et al., 2017). The limitation of the current controllable generation is no explicit control of other dimensions to maintain them the same. Our method addresses this issue by completely decomposing the input text into multiple labels with LLMs and then reconstructing it with switched attributes.

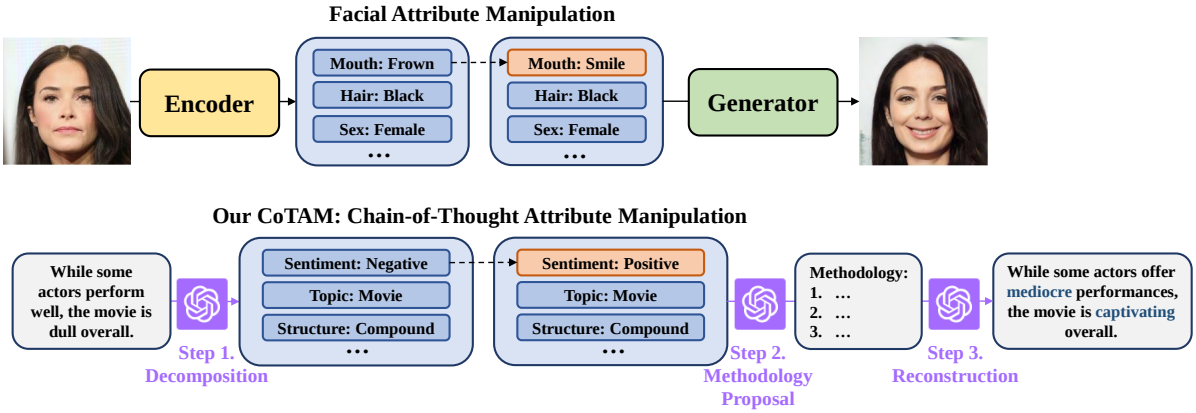

Figure 2: An overview of our CoTAM.

Large Language Models are large-scale models trained on a massive number of texts (Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022) that have been shown to have emerging capabilities (Wei et al., 2022b). One of these capabilities is learning from few-shot demonstrations, which is often referred to as in-context learning (Dong et al., 2022). However, these demonstrations must be concatenated into contexts during inference time, increasing the computational costs and carbon footprints. Another important capability is to follow instructions for zero-shot task transferrability (Wei et al., 2022a). Following this idea, ChatGPT (Ouyang et al., 2022; OpenAI, 2023) was trained with human feedback and reinforcement learning. Our work benefits from these instruction-tuned models to generate attributes and reconstruct sentences.

Data Augmentation is widely employed in low-resource scenarios to mitigate model overfitting. It is usually conducted in a label-preserving manner where only minor perturbations are added (Wei and Zou, 2019; Fadaee et al., 2017). Recently, a line of research propose to use LLMs for data augmentation. Specifically, they use few-shot data as demonstrations and prompt LLMs to generate new data (Yoo et al., 2021; Sahu et al., 2022a). They claim that the LLM is able to mix few-shot data and synthesize similar ones. Lin et al., 2023 further propose to use Pointwise V-information to filter unhelpful data from generations. Most recently Dai et al., 2023; Whitehouse et al., 2023 propose to generate data using ChatGPT and GPT-4 and observe performance improvement. Finally Cheng et al., 2023 use GPT-3 generated data to improve sentence embedding via contrastive learning. Our work aims at improving LLM-based data augmentation via attribute manipulation.

3 Our CoTAM

3.1 Preliminary

Language Modeling is the basis of the near-human language ability of LLMs. The training objective is to maximize the probability of the next token prediction $\prod_{i=1}^{n} p(w_n|\text{<sos>}, w_{1:n-1})$ on human texts. Here <sos> refers to the start-of-sequence token. By training LLMs on very huge corpora, current LLMs are able to follow human instructions to achieve outstanding zero-shot performance or process natural language. To instruct LLMs, one needs to input a prompt Z and the LLM will then generate the response $W \sim p(w_n|\text{<sos>}, Z, w_{1:n-1})$ based on it, which is then decoded to the output.[2] With CoT, the LLM is instructed to first solve simple prerequisite problems and then achieve the instruction goal better.

Fine-tuning is generally applied to train smaller models. The model has a text embedder $\mathcal{E}$ and a classifier $\mathcal{C}$. Taking an input text W, the embedder $\mathcal{E}$ first maps it to a representation $X \in \mathbb{R}^d$ where d is the latent space size. Then $\mathcal{C}$ further maps X to a probability distribution $P \in \mathbb{R}^c$ where c is the number of classes. The cross-entropy loss is then calculated between P and the ground truth Y, which is used to update the parameters in the model by backpropagation.

3.2 Framework of CoTAM

The aim of our CoTAM is to generate efficient data from LLMs that enable well-performed small mod-

[2] In our experiments, the response generally follows a decodable pattern when the temperature of the language model is set to near 0.

els with the least data for fine-tuning. Our idea is to generate pairs (groups) of data that are different in the classification target but the same in other attributes. In the context of fine-tuning smaller models, our method focuses on generating pairs or groups of data that have identical attributes, save for the feature used for classification. By making all attributes other than the target feature identical in each data pair or group, the changes in the resulting P across different pairs (groups) can be primarily attributed to variations in the target feature. This reduces the complexity typically encountered with noisy, real-world data. As a result, the cross-entropy loss between P and the ground truth Y, used to update the model parameters through backpropagation, becomes a more precise indicator of the target feature's impact on classification.

To create such data, we introduce attribute manipulation that is primarily applied for facial attribute manipulation. As shown in Figure 2, a learned encoder maps input images to representations in the latent space. Then, the representations are transformed in the latent space and reconstructed into images. Consequently, the reconstructed image will result in an explicit change in that attribute while preserving similar other attributes. Thus, the difference between the initial and the reconstructed image enables efficient training for the classifier on the switched attribute.

With the strong text manipulation capability of LLMs (OpenAI, 2023), we adapt attribute manipulation to texts via in-context interaction. To be more specific, we create CoT queries to decompose the input texts into many attributes, which approximates the latent space. The CoT then switches the target attribute in the task and prompts the LLM to reconstruct the manipulated sentence. The main challenge here is how to approximate the latent space by attributes. The latent space in facial attribute manipulation represents a set of fixed explicit or implicit attributes. However, the fixed attribute set is not applicable to texts as they generally do not share common attributes like facial images.

Encouragingly, the LLM has been demonstrated to proficiently propose textual attributes (Wang et al., 2023), meeting our requirement for dynamic attribute decomposition. This means that we can use a dynamic set attribute proposed by the LLM to represent different input texts. As such, we construct a CoT using LLMs, drawing inspiration from facial attribute manipulation. Firstly, the LLM is tasked with proposing several other attributes apart from the known one (the annotated label). Subsequently, the LLM is instructed to consider how to generate a sentence that **only** diverges from the switched label. Finally, the LLM is guided to compose such a sentence to finalize the attribute manipulation. In the CoT, the LLM, because of its powerful text manipulation capabilities, plays dual roles as both the decomposer and reconstructor.

3.3 Detailed CoT Implementation

What: Decomposition Following the macro-level design of CoTAM, the first step in the CoT is to decompose the sentence into various attributes. Our instruction in this step is

What are some other attributes of the above sentence except "<Attr>"?

where <*Attr*> here refers to the known attribute in the dataset. For instance, <*Attr*> can be *sentiment: positive* in a sentiment analysis task. Consequently, the LLM will propose a series of other attributes which should be maintained in the reconstruction.

How: Methodology In the second step, we will instruct the LLM to propose the methodology to reconstruct a sentence with the switched attribute and others from the decomposition step. This step is incorporated as understanding how to achieve the goal, which is important to the CoT inference (Wei et al., 2022d). Our instruction in this step is

How to write a similar sentence with these attributes and "<New Attr>"?

where <*New Attr*> is a switched <*Attr*> like *sentiment: negative* if <*Attr*> is *sentiment: positive*. This step will output a guideline for the LLM to finally execute the reconstruction.

Write: Reconstruction This step simply asks the LLM to reconstruct the sentence with one attribute switched using the following instruction,

Write such a sentence without any other explanation.

where the constraint "*without any other explanation*" is added only to improve the sampling efficiency.

4 Experiment

4.1 Datasets

We verify the advantage of CoTAM on text classification and other tasks using 6 datasets. 3 text

Dataset	Attributes
SST-2	sentiment: positive
TweetEmo	sentiment: anger
AG-News	topic: world news
MNLI	natural language inference: contradiction
MRPC	semantics: equivalent to sentence 1
CSQA	best choice: <answer name>

Table 1: Attribute name examples in datasets of our experiments.

classification datasets include SST-2 (sentiment polarity) (Socher et al., 2013), TweetEmo (fine-grained sentiment) (Barbieri et al., 2020), and AG-NEWS (topic) (Zhang et al., 2015). 3 other task datasets include MNLI (natural language inference) (Williams et al., 2018), MRPC (semantic textual similarity) (Dolan and Brockett, 2005), and CSQA (multiple choice question answering) (Talmor et al., 2019). MNLI includes matched (MNLI_m) and mismatched (MNLI_mm) datasets for evaluation. We report results on the validation dataset when the test dataset is not publicly available considering the efficiency to get multi-run results. The statistics of datasets are presented in Appendix A. We present some examples of attribute names in Table 1 and more details can be found in Appendix B.

4.2 Compared Methods[3]

Our CoTAM We query `GPT-4` (OpenAI, 2023) with the CoT to construct the dataset. The temperature of the LLM to set to 0 towards high quality and reproducibility. We apply CoTAM to 200 sentences in each dataset to create a small subset from which we sample training data. For fair comparison, this subset is also used in other baselines for data generation

CoT Data Augmentation (CoTDA) is a LLM-based augmentation strategy (Dai et al., 2023) refined by our CoT scenario. Instead of directly asking for augmentation, we let the LLM follow our proposed CoT and propose a methodology to write a sentence with the **same** attributes as the input sentence. CoTDA is the main baseline for comparison to explore the importance of attribute switching in our CoTAM. For each seed data, we augment it for N-1 times with 0.1 temperature, where N refers to the number of classes in the dataset. Thus, CoTDA generates the same number of new data as CoTAM to achieve a fair comparison.

[3]Prompts in our experiments are presented in Appendix C

FlipDA (Zhou et al., 2022) is a traditional label-switched augmentation method based on conditional generation by a fully-tuned T5 (Raffel et al., 2020). Specifically, the sentence is combined with the switched label as the input to T5. Then, some spans in the sentence are randomly masked and recovered by T5 conditioning on the new label to switch the semantics of the sentence. As the original FlipDA requires a large supervised dataset that is inapplicable to few-shot learning, we build an LLM-based FlipDA (FlipDA++) baseline by sending span replacement instructions to LLMs.

Human/LLM Annotation directly using the texts labeled by humans or LLMs. For human annotation, we include the K-shot and NK-shot setups. K-shot represents the baseline before integrating the data generated from LLMs. As NK-shot has the same number of training data as CoTAM but with human annotation, we expect NK-shot human annotation to be the upper bound of our method. Whereas, we will see CoTAM able to outperform this upper bound, which can be attributed to higher data quality resulting from attribute manipulation. NK-shot LLM annotation[4] represents a simple baseline that is generally applied when much unlabeled in-domain data is available.

By default, we set K to 10 and all reported results are the average over 10 runs to eliminate the bias.

4.3 Fine-tuning Result

A simple way to evaluate the data quality is to tune a model on it and then check its performance. We select `RoBERTa-Large` (Liu et al., 2019) as the learner on different datasets. With the validation dataset unavailable, we train the model for 32 epochs[5] and then evaluate it.

As presented in Table 2, our CoTAM achieves the best fine-tuning results on all 7 tasks in comparison with other LLM-based data generation methods. On most tasks, the two label-switching methods (FlipDA and CoTAM) outperform other methods, which indicates using the LLM to switch labels creates more efficient data. On label switching, attribute manipulation shows superiority over simple span replacement as our CoTAM performs better than FlipDA on all tasks. The prominent performance of CoTAM also verifies the capability of LLMs to manipulate complex attributes which

[4]K-shot data are used for in-context inference.

[5]Except 8 epochs for MRPC, on which the model is more likely to overfit.

Method	SST-2	TweetEmo	AG-NEWS	MNLI_m	MNLI_mm	MRPC	CSQA
$\text{K-Shot}_\text{Human}$	60.54	44.38	81.05	35.88	38.75	51.96	34.54
$\text{NK-Shot}_\text{Human}$	62.17	69.51	88.66	43.33	44.03	57.50	47.36
$\text{NK-Shot}_\text{LLM}$	61.14	69.11	85.64	41.71	42.92	55.88	45.12
K-FlipDA++	74.28	70.87	84.72	51.52	53.56	60.15	50.52
K-CoTDA	70.83	67.76	85.19	36.06	36.28	55.54	48.79
K-CoTAM	**79.12**	**72.76**	**85.80**	**54.07**	**56.16**	**61.64**	**53.22**

Table 2: Few-shot learning results based on data annotated by humans and LLMs. The data number used in each method is the same (N^2K) except the baseline $\text{K-Shot}_\text{Human}$ (NK). **Bold:** Best performance.

Method	SST-2		TweetEmo		AG-NEWS	
	NC	KNN	NC	KNN	NC	KNN
$\text{K-Shot}_\text{Human}$	82.00	78.20	66.01	59.92	77.72	73.57
$\text{NK-Shot}_\text{Human}$	87.55	83.45	71.23	67.56	84.70	82.33
$\text{NK-Shot}_\text{LLM}$	86.78	80.26	69.34	64.90	**81.19**	**79.34**
K-FlipDA++	88.13	86.76	66.53	65.05	79.82	75.11
K-CoTDA	86.38	83.00	68.63	61.58	78.87	76.56
K-CoTAM	**88.43**	**87.52**	**70.02**	**65.37**	80.60	75.48

Table 3: The performance with pre-trained text representation models. Few-shot results of instance-based algorithms on text classification.

Data	SST-2			MNLI
	T	NC	KNN	T
K-CoTAM	**79.12**	**88.43**	**87.52**	**54.07**
w/o What	75.69	88.03	86.78	45.61
w/o How	77.94	88.15	87.01	48.98
w/o CoT	71.82	87.94	86.24	39.34
w/ V3.5	72.93	87.59	84.31	41.32

Table 4: The ablation study on our CoTAM. Matched MNLI results are presented for analysis.

might refer to premises or questions.

On 6 out of 7 tasks, our CoTAM breaks the supposed upper boundary of (N-way) NK-shot human annotations. This indicates that carefully crafted data from LLMs have the potential to train better models than ones trained on the same number of human annotations. Also, aur CoTAM is verified to be such a method that improves the data efficiency by attribute manipulation.

4.4 Instance-based Algorithm Result

In the field of few-shot text classification, text embedding has proven to be a powerful tool for improving performance and efficiency (Muennighoff et al., 2023). This section is dedicated to exploring instance-based techniques designed explicitly for text classification with text embedding models.

In instance-based inference, a text embedding model converts the input sentence into a representation. The label of this representation is then determined based on its proximity to annotated sentence representations. We utilized two tuning-free algorithms in our experiments—Nearest Centroid (NC) (Manning et al., 2008) and K-Nearest Neighbors (KNN)—and applied them to three different text classification datasets. NC assigns a label to an input sentence depending on how close it is to centroids, defined as the average representation of sentences sharing the same label. In contrast, KNN labels the input sentence according to the most common label amongst its nearest K neighbors. We set K to 5 in our experiments. We harness the Simple Contrastive Sentence Embedding (SimCSE) model (Gao et al., 2021), with `RoBERTa-Large` as the backbone model[6], to encode the texts.

Table 3 showcases the performance of different data generation methods when used with instance-based algorithms. In contrast to methods that generate new texts (such as FlipDA and CoTDA), our proposed method, referred to as CoTAM hereafter, exhibits superior performance in most configurations. This implies that data created by CoTAM also benefits from improved distributions in the latent space of text embedding models. On the AG-NEWS dataset, instance-based algorithms show a preference for in-domain annotations, whether made by humans or Large Language Models (LLMs). This highlights the importance of using in-domain texts when employing these algorithms for certain tasks.

5 Further Analysis

5.1 Ablation Study

We launch an ablation study to verify the importance of each thought in the CoT. We also explore the effect of different LLMs. We thus change the LLM in our experiments to `GPT-3.5-turbo`. The experiments show that the `GPT-4` leads to significantly better fine-tuning results. Also, this gap can

[6] huggingface.co/princeton-nlp/sup-simcse-roberta-large

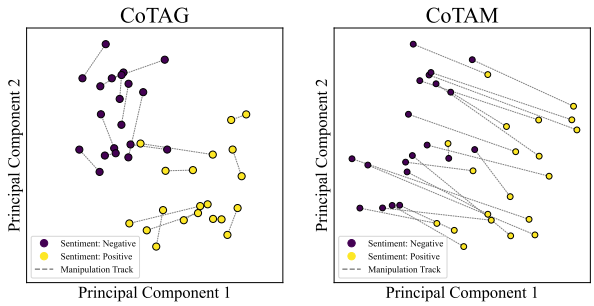

Figure 3: Principal component analysis of text pairs generated by our CoTAG and CoTAM on the SST-2 dataset.

be narrowed down by text embedding models on text classification.

The outcomes of our ablation study are detailed in Table 4. In this study, we found that eliminating each "thought" from our CoT resulted in a decline in performance. Interestingly, the "what" (decomposition) thought proved more critical than the "how" (methodology) thought, accentuating the predominance of attribute proposal over auxiliary methodology proposal. The CoT is necessary for label switching as the removal of it leads to significant performance degradation. Finally, `GPT-4` outperforms `GPT-3.5-turbo`, indicating that CoTAM favors larger LLM with better language capability, especially on more complex tasks like MNLI.

5.2 Visualization of Attribute Manipulation

In an attempt to confirm our hypothesis that LLM is adjusting a single feature while keeping other attributes constant, we illustrate data pair representations from CoTAM in Figure 3. We use principal component analysis (PCA) (F.R.S., 1901) to take the high-dimensional (1024-dimensional) text representations from SimCSE and simplify them into a 2-dimensional space for ease of visualization.

The diagram distinctly demarcates between positive and negative representations, which underscores the value of our method in fine-tuning and instance-based inference. Additionally, the direction of representation switching is largely consistent, providing further evidence that LLMs have the ability to tweak one attribute while keeping others stable. This consistency in the direction of the switch hints at the predictability and control we have exercised over LLM behavior for targeted feature manipulation. In comparison to CoTAG, our CoTAM depicts a clearer boundary, thus enabling more efficient data learning than traditional data augmentation.

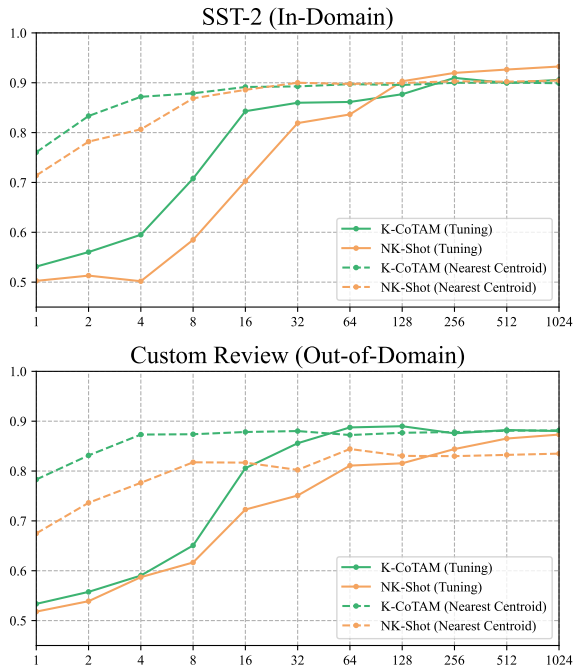

Figure 4: Performance comparison between K-CoTAM and NK-Shot on in-domain and out-of-domain test datasets.

5.3 Initial Data Analysis

In this section, we analyze how the performance of the CoTAM is affected by the initial data (N-way K-shot input). We also utilize the property of CoTAM to handle scenarios with less human participation.

Data Scale We first analyze how the number of initial data affects the performance of our CoTAM. Thus, we sample 3000 more instances from SST-2 to scale up the sampling pool. As presented in Figure 4, K-CoTAM is able to break the NK-Shot boundary with few examples ($K \leq 64$) for fine-tuning. With text representation models, K-CoTAM shows a significant advantage on very few examples ($K \leq 4$) and converges to a similar performance with human annotation. Though fine-tuning on more human annotation leads to higher performance than CoTAM, the in-domain performance improvement might be a result of overfitting to the domain. Thus, we further evaluate K-CoTAM and NK-Shot on custom review, an out-of-domain dataset with the same labels as SST-2. On custom review, K-CoTAM shows a consistent advantage with different data numbers. Thus, we conclude our CoTAM is more robust to domain mismatching than direct tuning.

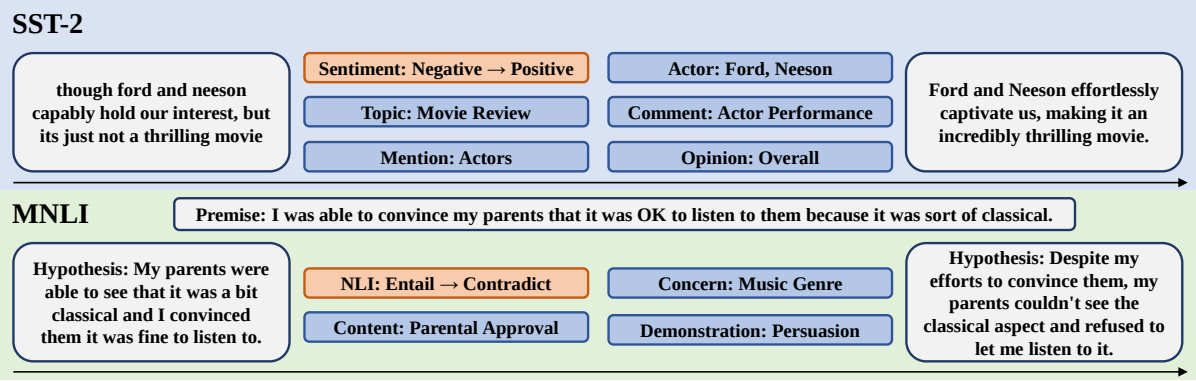

Figure 5: Case study of the real workflow in CoTAM.

Data	SST-2			MNLI
	T	NC	KNN	T
K-CoTAM	79.12	88.43	87.52	54.07
1-way K-shot	53.56	84.03	76.72	39.88
1-way NK-shot	58.47	84.51	81.50	47.27
LLM-proposed	85.81	86.23	85.16	51.95

Table 5: The initial data analysis of our CoTAM.

Single Label Data One notable advantage of the CoTAM approach is its potential to generate data labels that were not included in the original training dataset. This unique feature enables us to simplify the annotation process by reducing multiple categories to just a single one. In this context, we examined two setups: 1-way K-shot and 1-way NK-shot data. As displayed in Table 5, the shift to single-label setups does not lead to a substantial drop in performance, with the exception of fine-tuning on SST-2. This suggests the challenges associated with parameter tuning in the absence of in-domain data for each label, as the performance on instance-based algorithms is not markedly impacted.

LLM-Proposed Data approach takes us a step closer to the near elimination of human annotators. In this setup, we employ the LLM to generate the preliminary data and subsequently invert the sentences. As depicted in Table 5, the LLM-proposed data achieves a performance level comparable to that of human annotations. This underscores the potential of LLMs to almost completely do away with human involvement while efficiently training smaller models.

5.4 Case Study

Figure 5 specifies the real attribute manipulation process in our experiments. For better depiction, we simplify the response by only presenting the attributes proposed by the LLMs.

In the SST-2 example, other attributes include labels in a different categorization (Topic: Movie Review), actor entities (Actor: Ford, Neeson), and overall style (Opinion: Overall). These attributes are well preserved in the reconstruction, which contributes to a strong contrast in the task target and consequently improves the data efficiency.

Moving on to the MNLI example, the sentence primarily breaks down into different semantic elements. When these elements are reconstructed, they follow a logical sequence that differs from the original sentence. Thus data from CoTAM reinforces the learner's comprehension of textual logic which is crucial for tackling MNLI.

6 Conclusion

The study introduces a novel method, Chain-of-Thoughts Attribute Manipulation (CoTAM), which uses manipulated data from Large Language Models (LLMs) for few-shot learning. CoTAM, inspired by image manipulation, creates label-switched data by modifying task-specific attributes and reconstructing new sentences. Our testing validated the effectiveness of CoTAM over other LLM-based text generation techniques. The results also showcase the potential for LLM-guided learning with less supervision.

Future work will aim to adapt the attribute manipulation technique for smaller language models, increasing its scalability and accessibility. This would reduce reliance on the resource-intensive processes inherent to large language models, improving efficiency. Additionally, efforts will be made to ensure output stability and reduced supervision, making CoTAM practical for real-time applications while preserving performance quality.

Limitation

Despite the significant advancements in few-shot learning and attribute manipulation reported in this paper, our proposed CoTAM does come with certain limitations. Firstly, our approach leverages a chain-of-thoughts decomposition and reconstruction procedure which, while yielding improved data efficiency and model performance, tends to result in a decrease in the overall generation efficiency compared to traditional methods. This may affect the method's scalability, particularly in scenarios requiring rapid data generation. Secondly, the current implementation of CoTAM is primarily confined to attribute-related tasks, limiting its scope of application. While this constraint is a direct result of our method's design focused on manipulating task-specific attributes, we acknowledge that extending CoTAM's applicability to a broader set of tasks could significantly increase its utility. Our future work will thus aim to address this limitation. Lastly, it should be noted that the effectiveness of CoTAM is fundamentally dependent on the abilities of the underlying Large Language Models. As a consequence, the limitations inherent in these LLMs, such as biases in their training data or limitations in their understanding of nuanced contexts, could potentially impact the performance of CoTAM. It is thus crucial to continually improve and refine the LLMs used in our method to ensure the accuracy and robustness of the generated data.

Ethical Consideration

Our work instructs large language models to generate efficient training data, which generally does not raise ethical concerns.

References

Francesco Barbieri, José Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1644–1650. Association for Computational Linguistics.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 10–21. ACL.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Qinyuan Cheng, Xiaogui Yang, Tianxiang Sun, Linyang Li, and Xipeng Qiu. 2023. Improving contrastive learning of sentence embeddings from ai feedback. *arXiv preprint arXiv:2305.01918*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Tianming Liu, and Xiang Li. 2023. Chataug: Leveraging chatgpt for text data augmentation.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing, IWP@IJCNLP 2005, Jeju Island, Korea, October 2005, 2005*. Asian Federation of Natural Language Processing.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Karl Pearson F.R.S. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals,

and L. Sifre. 2022. Training compute-optimal large language models. *ArXiv*, abs/2203.15556.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of text. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1587–1596. PMLR.

Canasai Kruengkrai. 2019a. Learning to flip the sentiment of reviews from non-parallel corpora. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6311–6316, Hong Kong, China. Association for Computational Linguistics.

Canasai Kruengkrai. 2019b. Learning to flip the sentiment of reviews from non-parallel corpora. 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 6310–6315. Association for Computational Linguistics.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and José M. Álvarez. 2016. Invertible conditional gans for image editing. *CoRR*, abs/1611.06355.

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

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022a. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam H Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022b. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9240–9249. Computer Vision Foundation / IEEE.

Yujun Shen and Bolei Zhou. 2021. Closed-form factorization of latent semantics in gans. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1532–1540. Computer Vision Foundation / IEEE.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4149–4158. Association for Computational Linguistics.

Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-driven explainable clustering via language descriptions. *CoRR*, abs/2305.13749.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022b. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022c. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022d. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Jason 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)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. Llm-powered data augmentation for enhanced crosslingual performance. *arXiv preprint arXiv:2305.14288*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Taihong Xiao, Jiapeng Hong, and Jinwen Ma. 2018. ELEGANT: exchanging latent encodings with GAN for transferring multiple face attributes. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part X*, volume 11214 of *Lecture Notes in Computer Science*, pages 172–187. Springer.

Kevin Yang and Dan Klein. 2021. FUDGE: controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3511–3535. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

Jing Zhou, Yanan Zheng, Jie Tang, Li Jian, and Zhilin Yang. 2022. Flipda: Effective and robust data augmentation for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8646–8665. Association for Computational Linguistics.

A Dataset Statistics

	Dataset	SST-2	TweetEmo	AG-News
TC	Domain	Sentiment	Sentiment	Topic
	#Test	$1.8K$	$1.4K$	$7.6K$
	#Label	2	4	4
	Dataset	MNLI	MRPC	CSQA
Others	Task	NLI	STS	MCQA
	#Test	$9.8K$	$1.7K$	$1.1K$
	#Label	3	2	5

Table 6: The statistics of datasets in our experiments.

The statistics of the dataset used in the experiments are presented in Table 6. The numbers of test instances in matched and mismatched are both $9.8K$.

B Attribute Names

Dataset	Attributes
SST-2	sentiment: positive sentiment: negative
TweetEmo	sentiment: anger sentiment: joy sentiment: optimism sentiment: sadness
AG-News	topic: world news topic: sports news topic: business news topic: sci/tech news
MNLI	natural language inference: contradiction natural language inference: neutral natural language inference: entailment
MRPC	semantics: equivalent to sentence 1 semantics: inequivalent to sentence 1
CSQA	best choice: <answer name>

Table 7: The attribute names in datasets of our experiments.

The attribute names of the dataset used in the experiments are presented in Table 7.

C Prompts

Target	Prompt
CoTAM	"<sentence>" Please think step by step: 1. What are some other attributes of the above sentence except "<attr>"? 2. How to write a similar sentence with these attributes and "<new attr>"? 3. Write such a sentence without any other explanation.
CoTAG	"<sentence>" Please think step by step: 1. What are some other attributes of the above sentence except "<attr>"? 2. How to write a similar sentence with these attributes and "<attr>"? 3. Write such a sentence without any other explanation.
FlipDA	"<sentence>" Please think step by step: 1. How to switch the above sentence to "<new attr>" by changing some spans? 2. Write the switched sentence without any other explanation.

Table 8: The prompts used in our experiments.

The prompts used in the experiments are presented in Table 8.