Controllable Data Augmentation for Few-Shot Text Mining with Chain-of-Thought Attribute Manipulation

Anonymous ACL submission

Abstract

Prompting large language models (LLMs) for data augmentation has recently become a common practice in few-shot NLP tasks. In this paper, we propose Chain-of-Thought Attribute Manipulation (CoTAM), a novel approach that generates new data from existing examples by only tweaking in the user-provided, task-specific attribute, e.g., sentiment polarity or topic in movie reviews. Instead of conventional latent representation controlling, we leverage the chain-of-thought prompting to directly edit the text in three steps, (1) attribute decomposition, (2) manipulation proposal, and (3) sentence reconstruction. Extensive results on various tasks, such as text (pair) classification and aspect-based sentiment analysis, verify the superiority of CoTAM over other LLM-based augmentation methods with the same number of training examples for both fine-tuning and in-context learning. Remarkably, the 2D visualization of the augmented dataset using principle component analysis revealed a human-recognizable decision boundary that is likely hinted by the attribute manipulation, demonstrating the potential of our proposed approach.

1 Introduction

Prompting large language models (LLMs) for data augmentation has recently become a common practice in few-shot natural language processing (NLP) tasks. Existing methods (Yoo et al., 2021; Sahu et al., 2022b; Dai et al., 2023; Lin et al., 2023) typically first generate new task-specific data with LLMs hinted by few-shot demonstrations, and then fine-tune a (small) pre-trained language model with the augmented dataset for better performance. The same augmented data can be also incorporated into in-context learning (ICL) (Li et al., 2023; Dong et al., 2023). However, these augmentation methods usually prompt LLMs to generate new examples *wildly* without proper control, which hinders the informativeness of generated data and might induce spurious correlation. As shown Figure 1(left), the generated data without control has no clear pattern and could even possibly mislead the fine-tuning or ICL under few-shot learning scenarios.

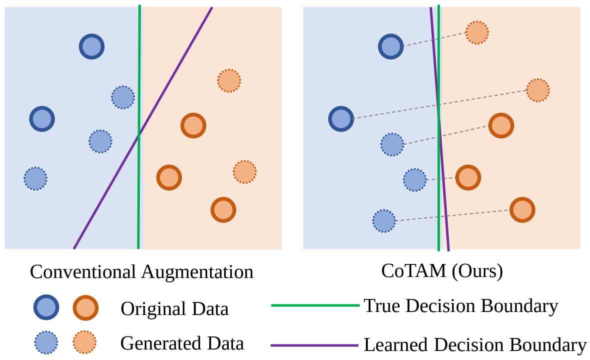

Figure 1: An illustrative comparison in case of binary classification. Conventional data augmentation generates uncontrolled data, while CoTAM directly reflects decision boundaries through task instructions. We present a real example in Figure 4.

In this paper, we propose a controllable data augmentation for few-shot text mining. The general idea is to generate new data from existing examples by only tweaking in the user-provided, task-specific attribute, e.g., sentiment polarity or topic in movie reviews. Intuitively, as shown in Figure 1, one can expect that this approach can efficiently find the decision boundary because we (1) directly manipulate along the direction of task-specific attributes and (2) maintain the rest of the attributes as before.

Different from the existing controllable generation works in computer vision (Shen et al., 2020; Shen and Zhou, 2021) and natural language processing (Kruengkrai, 2019a; Zhou et al., 2022), where attributes are manipulated in the latent space of the encoder before reconstructing new instances, we leverage the chain-of-thought (CoT) prompting (Wei et al., 2022c) to directly edit the text using LLMs in three steps, (1) attribute decomposition, (2) manipulation proposal, and (3) sentence reconstruction. Specifically, we start with the user-provided, task-specific attributes, and then prompt

LLMs to decompose each individual text example into other orthogonal attributes. Compared with a pre-defined attribute set per dataset, we believe that such dynamically constructed, per-example sets of attributes can better capture the uniqueness of every piece of text. Second, we instruct LLMs to propose a plan to manipulate the values of the task-specific attributes while maintaining the other attribute values the same. Finally, we prompt the LLMs to reconstruct the sentence based on the manipulation proposal. All these steps are written in a single prompt and fed to the LLM at once. Furthermore, using LLMs benefits the interpretability of our framework where attributes are completely transparent to users.

We conduct extensive experiments to evaluate CoTAM and baselines using a series of few-shot classification tasks with very different classification targets and aspect-based sentiment analysis for more complex attribute manipulation. For fair comparison, all compared methods utilize the same LLMs and generate the same amount of data. We assess the quality of generated data by looking at (1) the performance of trained small language models via fine-tuning or tuning-free methods on the augmented data and (2) the ICL performance of LLMs using the augmented data as demonstrations. Extensive experimental results including label-scarce and out-of-domain scenarios demonstrate the advantage of proposed controllable data augmentation over conventional methods. The ablation study further reveals the necessity of attribute manipulation comparing to directly flipping the labels. Finally, we present PCA analysis on the embeddings of generated augmentations that visually illustrates the effectiveness of method.

Our contributions are three-fold:

- We propose a novel controllable data augmentation approach CoTAM based on chain-of-thoughts prompting using LLMs, which directly edits the text examples in an interpretable way instead of tweak latent representation vectors.
- We conduct experiments on a wide spectrum of tasks and datasets, demonstrating the effectiveness of the augmented data by CoTAM in both fine-tuning and in-context learning.
- Our detailed analyses, especially the human-recognizable decision boundaries revealed by the 2D visualization of the augmented dataset using principle component analysis, demonstrate the significant potential of our proposed attribute manipulation approach.

Reproducibility. We will open-source the code. [1]

2 Problem Formulation

We aim to generate more efficient training data using controllable augmentation on a few-shot dataset $\mathcal{D}$ focusing on a **target attribute** Y (e.g., the classification objective) with N possible values $\{y_1, y_2, \cdots, y_N\}$ (i.e., N**-way**). For each possible attribute value y_i, the dataset $\mathcal{D}$ provides K examples (i.e., K**-shot**) of texts with the value. We here showcase two mainstream few-shot learning schemes as the basis to discuss the augmentation:

- **In-context Learning** (ICL) is a scheme for LLMs, which takes a few examples of sentences with their target attribute values (i.e., a series of (X, y_i)) as the context to handle new inputs. With these demonstrations, the LLM is expected to understand the underlying mapping and then predict the label of new inputs.
- **Fine-tuning** generally trains smaller models with the (limited) labeled data. The model has a text embedder $\mathbf{E}$ and a classifier $\mathbf{C}$. A text x from the dataset $\mathcal{D}$ will be represented as a dense vector $\mathbf{E}(x)$, which is learned to encode the attributes of x, including the target attribute Y and other attributes. The classifier $\mathbf{C}$ further processes the vector $\mathbf{E}(x)$ and outputs a distribution over $y_1, y_2, \cdots, y_N$, indicating the probability of each Y value in x.

Ideally, our controllable augmentation shall supply efficient demonstrations and training data under the ICL and fine-tuning settings, respectively.

3 Our CoTAM Framework

To boost the performance of few-shot methods, we suppose a scenario, shown in Figure 2, to augment examples that well improve the task awareness of the inference models. For a given sample x with target attribute value y from $\mathcal{D}$, we will manipulate its attribute value to y' that $y \neq y'$ to form a build a new sentence x'. We set two requirements for the manipulation: 1) Significant Manipulation on the target attribute Y, which means the manipulated result x' should be viewed with y_j by oracle like humans. 2) Minor Manipulation on all other attributes $\mathcal{Z}$, which indicates x and x' to share a similar value z_k for all $Z \in \mathcal{Z}$. To meet the two requirements above will ensure x and x' only differ in attribute Y, making them an efficient pair

[1] Code: https://github.com/anonymous_repo

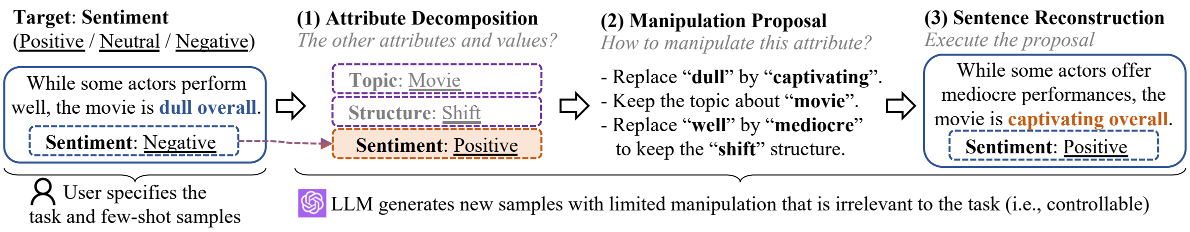

Figure 2: An overview of the goal and implementation of our CoTAM.

for learning on the dataset $\mathcal{D}$. Take fine-tuning as an example, the loss $\mathcal{L}(X, y_i) + \mathcal{L}(X', y_j)$ will be attributed to the only different attribute Y, thus let each annotation by humans efficiently reflect the target attribute with its augmentations.

Based on our desiderata above, we propose CoTAM that benefits from the strong text manipulation capability of LLMs (OpenAI, 2023) with its workflow demonstrated in Figure 2. To be more specific, we first create chain-of-thought (CoT) queries to decompose the input texts into many attributes, which approximates the latent space. We aim to get rid of human labor to propose other possible attributes for efficiency. Moreover, in some cases, even human experts cannot give you a complete list of other attributes among all the possible texts. Finding a shared and fixed set of attributes for various kinds of texts is hard since different sentences rarely share a common set of applicable attributes. Encouraged by Wang et al. (2023), we instruct LLMs to propose a dynamic attribute set for each input text, which are customized among inputs dependent on which attributes are applicable. The CoT then switches the value of the target attribute to other possible values in the task and prompts the LLM to reconstruct the manipulated sentence. Finally, the LLM is guided to compose such a sentence to finalize the attribute manipulation.

Different from the existing controllable generation works in computer vision (Shen et al., 2020; Shen and Zhou, 2021) and natural language processing (Kruengkrai, 2019a; Zhou et al., 2022), where attributes are manipulated in the latent space of the encoder before reconstructing new instances, our CoTAM is proposed to directly edit the text using LLMs.

3.1 Step 1: Attribute Decomposition

Following the macro-level design of CoTAM, the first step in the CoT is to decompose the sentence into various attributes. The LLM takes the sentence and a human-annotated attribute-value pair as the input and then propose other attributes and their values.

For example, The sentence "*While some actors perform well, the movie is dull overall*" with be processed with its known attribute-value y_i, here is "*Sentiment: Negative*". The LLM then proposes a set of other applicable attribute-values $\hat{\mathcal{Z}} = \text{LLM}_{\text{AD}}(X, y_i) \subset \mathcal{Z}$ like "*Topic: Moive*", "*Structure: Shift*" as in Figure 2, which is a subset of $\mathcal{Z}$ but is generally detailed enough to approximate the irrelevant attributes. The value of the known attribute is then flipped to another one given by the user like "*Sentiment: Positive*", which is then combined with other LLM-proposed attribute-values for the next step.

3.2 Step 2: Manipulation Proposal

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., 2022c). In this step, the LLM takes all elements in the manipulation as the input and proposes an instruction $I = \text{LLM}_{\text{MP}}(X, y_i, y_j, \hat{\mathcal{Z}})$ for LLM to execute in the next step. A proposed manipulation is shown as in Figure 2, the LLM suggest several instructions to complete the manipulation.

3.3 Step 3: Sentence Reconstruction

This step simply asks the LLM to follow its proposed manipulation instruction I to reconstruct the sentence and output a label-flipped one as $X' = \text{LLM}_{\text{SR}}(X, I)$. As in Figure 2, the LLM follows the self-generated instructions to edit the input sentence to generate our desired X' that has significant different in Y (sentiment polarity) and minor difference in $\hat{Z}$ (proposed other attributes).

Dataset	Target Attribute	Possible Value
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>
ABSA	Sentiment on <Aspect>	Positive

Table 1: Target attributes and possible values in datasets of our experiments and more details can be found in Appendix B.

4 Experiments

In this section, we evaluate different LLM-based augmentation methods on a series of classification tasks, with different target attributes. We incorporate comprehensive ways of utilizing augmentations with different classification techniques, such as fine-tuning, in-context learning and inference with sentence embedding. We further evaluate the augmentation ability of methods on more complex tasks like aspect-based sentiment analysis.

4.1 Datasets

We verify the advantage of CoTAM on text classification and other tasks using 6 classification datasets, including **SST-2** (sentiment polarity) (Socher et al., 2013), **TweetEmo** (fine-grained sentiment) (Barbieri et al., 2020), AG-NEWS (topic) (Zhang et al., 2015), **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 ($\mathbf{MNLI_m}$) and mismatched ($\mathbf{MNLI_{mm}}$) datasets for evaluation. To further test the ability of CoTAM on attributes other than classification targets, we include a manipulation on aspect-based sentiment analysis (ABSA) datasets, **Restaurant** and **Laptop** from SemEval2014 (Pontiki et al., 2014). For ICL, we report the results on 1000 samples from the mixture of validation and test dataset due to cost issue. For other setups, 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.

4.2 Compared Methods

CoT Data Augmentation (CoTDA) is a augmentation variant of our method which applies a similar CoT for conventional augmentation. 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.

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 (**Base**) and NK-shot (**Extra Annotation**) setups. K-shot represents the baseline before integrating the data generated from LLMs. NK-shot has the number of training data after augmentation with human annotation, thus we expect it to be a upper bound of augmentation methods. 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[2] (**Pseudo Label**) represents a simple baseline that is generally applied when much unlabeled in-domain data is available.

Comparison Fairness We select `GPT-4` (OpenAI, 2023) as the LLM to construct the dataset. The temperature of GPT-4 to set to 0 towards high quality and reproducibility. We apply each augmentation method to a fixed subset of 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. By default, we set K to 10 for fine-tuning and 3 to ICL. All reported results are the average over 10 runs

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

	Method	SST-2	TweetEmo	AG-NEWS	$MNLI_m$	$MNLI_{mm}$	MRPC	CSQA
Fine-tuning	Base	60.54	44.38	81.05	35.88	38.75	51.96	34.54
	Extra Annotation†	62.17	69.51	88.66	43.33	44.03	57.50	47.36
	LLM Pseudo Label	61.14	69.11	85.64	41.71	42.92	55.88	45.12
	FlipDA++	74.28	70.87	84.72	51.52	53.56	60.15	50.52
	CoTDA	70.83	67.76	85.19	36.06	36.28	55.54	48.79
	CoTAM	**79.12**	**72.76**	**85.80**	**54.07**	**56.16**	**65.83**	**53.22**
ICL	No Example	90.50	69.80	81.30	67.50	**69.70**	69.80	73.50
	Base	94.00	74.50	85.50	68.10	68.10	70.60	76.30
	Extra Annotation†	94.70	79.00	88.70	68.70	68.60	71.40	76.80
	LLM Pseudo Label	94.20	75.80	85.80	66.90	69.00	67.90	76.50
	FlipDA++	94.30	76.70	85.20	68.80	68.90	70.70	77.00
	CoTDA	94.00	76.50	86.00	68.20	68.50	70.00	76.70
	CoTAM	**94.50**	**77.10**	**86.40**	**69.70**	69.20	**70.90**	**77.30**

Table 2: Few-shot learning results based on data annotated by humans and LLMs. †: Extra Annotation increases the number (NK) of human-annotated samples to the same number as LLM-annotated to compare the annotation ability between LLMs and humans. **Bold:** The best result with the base number (K) of human annotation, thus excluding "Extra Annotation".

Method	SST-2		TweetEmo		AG-NEWS	
	NC	KNN	NC	KNN	NC	KNN
Base	82.00	78.20	66.01	59.92	77.72	73.57
Extra†	87.55	83.45	71.23	67.56	84.70	82.33
LLM SL	86.78	80.26	69.34	64.90	**81.19**	**79.34**
FlipDA++	88.13	86.76	66.53	65.05	79.82	75.11
CoTDA	86.38	83.00	68.63	61.58	78.87	76.56
CoTAM	**88.43**	**87.52**	**70.02**	**65.37**	80.60	75.48

Table 3: Utilization of sentence embeddings for classification tasks based on different augmented few-shot examples.

(except for ICL due to expense) to eliminate the bias.

All the prompts in our experiments are presented in Appendix C for better reproducibility.

4.3 Classification Result

Fine-tuning 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[3] 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 might refer to premises or questions.

On 6 out of 7 tasks, our CoTAM breaks the supposed upper boundary of (N-way) NK-shot with extra 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.

In-context Learning The performances of ICL-based inference with different augmentation methods are demonstrated in Table 2. Our CoTAM show superior ability on providing LLMs with few-shot examples for inference, thus broadening the application of our method. The only fail case for CoTAM is the out-of-domain MNLI, where few-shot examples do not benefit the inference. Still, among all augmentation scenarios, our CoTAM performs the best for this evaluation.

Inference w/ Sentence Embedding 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 sen-

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

Method	Restaurant			Laptop		
	P.	R.	F.	P.	R.	F.
Base	30.61	40.38	34.82	23.73	28.57	25.93
Extra†	54.70	66.67	60.09	59.18	44.62	50.88
LLM SL	44.26	56.25	49.54	18.56	22.73	14.09
FlipDA++	45.90	58.33	51.38	26.58	42.86	32.81
CoTDA	44.55	51.04	47.57	26.09	36.74	30.51
CoTAM	50.00	64.58	56.36	33.33	44.90	38.26

Table 4: The performance of span manipulation on aspect-based sentiment analysis datasets.

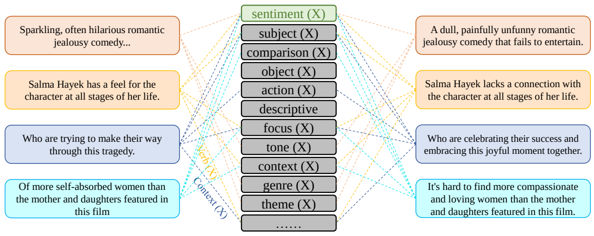

Figure 3: The workflows of CoTAM for different inputs.

tence 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[4], 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.

4.4 Aspect-based Sentiment Analysis

Here we further expand the utility of CoTAM to a more complex scenario to manipulate multiple span representations. We experiment on aspect-based sentiment analysis (ABSA), which aims to extract spans targeted by sentiment (aspects) in a statement and corresponding polarities. For instance, the aspect extracted from "The food is good." will be "positive aspect: food".

For attribute manipulation on ABSA, we view the aspects as the ABSA attributes like "positive aspect: food". We query the LLMs to decompose texts into ABSA and other attributes. The polarities of ABSA attributes are then randomly switched and used for the reconstruction. The reconstructed data are merged into the initial dataset as the augmentation.

We use the SemEval2014 ABSA dataset which has two subsets: restaurant and laptop and three sentiment polarities: positive, negative, and neutral[5]. We set the shot number (K) to 10 and generate 2 times for each instance ($N = 3$), which is the maximal manipulation time for an instance with only one aspect. The results on ABSA are presented in Table 4, our CoTAM successfully outperforms other LLM-based augmentation methods, which confirms that CoTAM is applicable to more complex scenarios than single sentence attribute manipulation.

5 Further Analysis

5.1 Workflow Demonstration

In Figure 3, we demonstrates the workflow of the dynamic attribute decomposition mechanism. In the workflow, our CoTAM decomposes sentences into applicable attributes and reconstructs while maintaining these attributes. For instance, *tone (X)* is more applicable to the first sentence due to its subjectivity and *comparison (X)* is more applicable to the last sentence since only it involves comparison. These attributes comprehend the unchanged parts of texts to guide the reconstruction during the manipulation. Subsequently, the reconstruction switch the targeted label (*sentiment (X)* in the case) with minor change to other attributes.

5.2 Ablation Study

We launch an ablation study to verify the importance of each thought in the CoT. We also explore

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

[5]We remove the conflict polarity because of its sparsity in the dataset.

Data	SST-2			MNLI
	T	NC	KNN	T
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
w/ FAP	76.38	87.79	85.13	47.91

Table 5: The ablation study on our CoTAM. Matched MNLI results are presented for analysis.

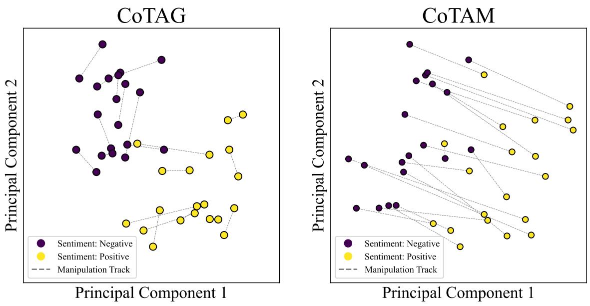

Figure 4: Principal component analysis of text pairs generated by our CoTDA and CoTAM on the SST-2 dataset.

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 be narrowed down by text embedding models on text classification.

The outcomes of our ablation study are detailed in Table 5. 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. In comparison between LLMs, `GPT-4` outperforms `GPT-3.5-turbo`, indicating that CoTAM favors larger LLM with better language capability, especially on more complex tasks like MNLI. Finally, we compare the performance of between CoTAM with a fixed attribute pool (FAP) and with a dynamic attribute pool in our experiments. The result shows the advantage to remove the type limitation of attribute the LLM decomposes into.

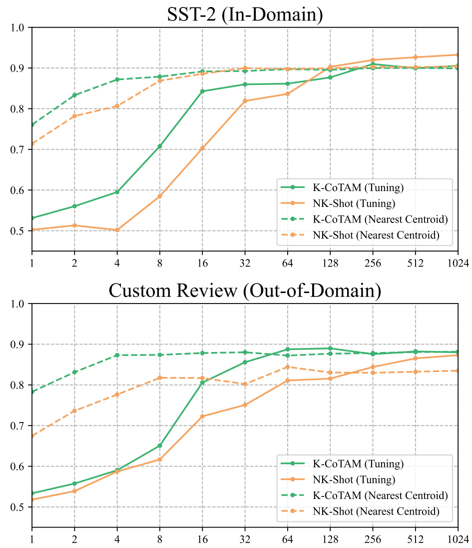

Figure 5: Comparison between K-shot CoTAM and NK-shot on in-domain and out-of-domain test datasets.

5.3 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 4. 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 CoTDA, our CoTAM depicts a clearer boundary, thus enabling more efficient data learning than traditional data augmentation.

5.4 Data Scale Analysis

In this section, we 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 5, CoTAM is able to break the NK-Shot boundary with few examples ($K \leq 64$) for fine-tuning. With text representation models, 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 CoTAM and NK-Shot on custom review, an out-of-domain dataset with the same labels as SST-2. On custom review, CoTAM shows a consistent advantage with different data numbers. Thus, we conclude our CoTAM is more robust to domain mismatching than direct tuning.

6 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.

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.

7 Conclusion

The study introduces a novel method, Chain-of-Thought Attribute Manipulation (CoTAM), which uses manipulated data from Large Language Models (LLMs) for few-shot learning. Our CoTAM 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.

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, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

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.

Dawei Li, Yaxuan Li, Dheeraj Mekala, Shuyao Li, Yulin wang, Xueqi Wang, William Hogan, and Jingbo Shang. 2023. Dail: Data augmentation for in-context learning via self-paraphrase.

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.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*, pages 27–35. The Association for Computer Linguistics.

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, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. 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.
CoTDA	"<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.

D Case Study

Figure 6 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 el-

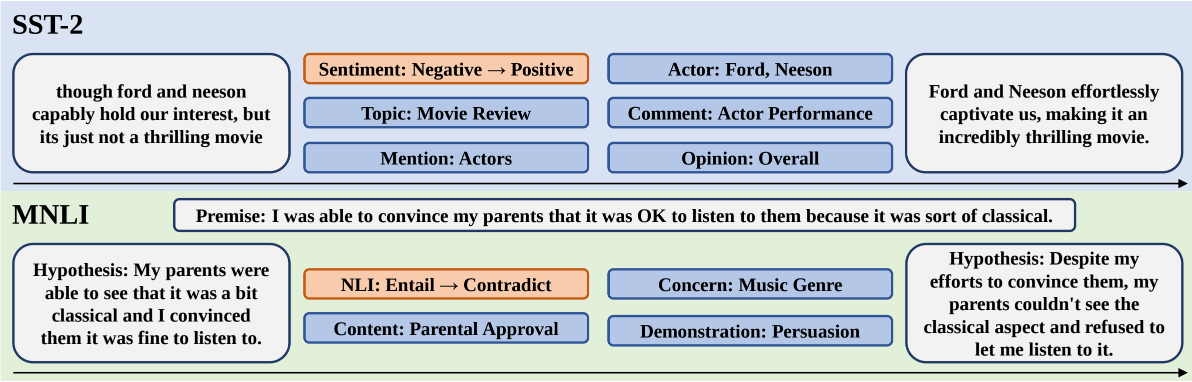

Figure 6: Case study of the real workflow in CoTAM.

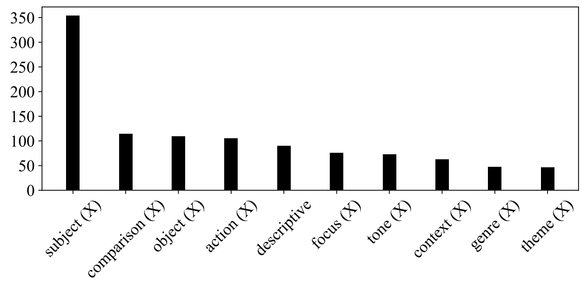

Figure 7: The statistics the most frequent 10 attributes in the decomposition step of CoTAM.

ements. 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.

E Attribute Statistics

In this section, we further explore the dynamic attribute decomposition mechanism in CoTAM. For 1315 instances from SST-2, there are 4513 decomposed attributes (3.43 per instance) and 2409 different ones. The distribution is in a long-tail pattern with 2124 attributes only appearing once. We show the statistics the most frequent 10 attributes from the decomposition in Table 7. We can observe a semantic diversity among the attributes, which verifies the ability of LLMs to comprehend the features of different inputs. As the most popular attribute *subject (X)* only appear in about 20%, there is no dominant attribute in the decomposition, which shows the flexibility of LLM-driven feature analysis. We also provide a quantitative comparison with a fixed feature pool in the ablation study.