Output Composability of QLoRA PEFT Modules for Plug-and-Play Attribute-Controlled Text Generation

Anonymous ACL submission

Abstract

Parameter-efficient fine-tuning (PEFT) techniques offer task-specific fine-tuning at a fraction of the cost of full fine-tuning, but require separate fine-tuning for every new task (combination). In this paper, we explore three ways of generalising beyond single-task training/inference: (i) training on combinations of multiple, related datasets; (ii) at inference, composing the *weight matrices* of separately trained PEFT modules; and (iii) at inference, composing the *outputs* of separately trained PEFT modules. We test these approaches on three different LLMs, QLoRA as the PEFT technique, and three sets of controlled text generation datasets for sentiment control, topic control, and multi-attribute control. We find that summing PEFT module outputs is a particularly strong composition method, which consistently either outperforms or matches the performance of alternative approaches. This is the case even when comparing against single-task specialised modules on the single-task test set, where three-module output composition achieves an average 2% point performance *increase* across all models for sentiment control.

1 Introduction

Given the substantial costs of training state-of-the-art language models, parameter-efficient finetuning (PEFT) techniques such as Adapters (Houlsby et al., 2019), Prefix Tuning (Li and Liang, 2021), and LoRA (Hu et al., 2021) are have become an important part of the toolbox for task-specific adaptation of pretrained models. PEFT techniques produce parameter matrices that are modular in the sense that they can be attached, detached and replaced in the same host model as needed, and this holds out the appealing (but currently far from realisable) vision of full plug-and-playability where individual task-specifically trained modules can be combined at will to achieve specific model behaviours, potentially even in conjunction with different host models (Sabry and Belz, 2024).

One aim in combining modules in this way is to achieve the combined functionality, e.g. controlling multiple attributes of generated text for which the individual modules have been (separately) trained. However, it is also desirable to preserve performance at the separate tasks. We examine the extent to which (i) composite-task performance can be achieved, and (ii) single-task performance preserved, when QLoRA PEFT modules are combined in host models. We test the **composability** of PEFT modules in this sense with different composition techniques on the composite task of multiple-attribute control, and the single tasks of sentiment and topic control. We compare output and parameter composition, as well as a baseline of training a single module on multiple datasets.

We focus on attribute-controlled text generation tasks, because these well established tasks enable systematic evaluation across multiple datasets while providing a controlled setting to isolate the core mechanics of PEFT module composition.

In presentation order, our main contributions are:

1. A new plug-and-play implementation of QLoRA supporting PEFT module composition for any number of modules, with three different composition techniques (Section 3);
2. New output composition techniques for PEFT modules (Section 3.1);
3. A new method for combining disparate datasets into a single, representative dataset (Section A); and
4. New test results for all composition techniques on 3 models, 14 datasets, and 3 composition techniques (Section 7).

2 Related Research

Parameter-efficient fine-tuning (PEFT) has been shown (Liu et al., 2024; Hu et al., 2021; Poth et al., 2023; Whitehouse et al., 2024) to effectively inject

task-specific knowledge into pretrained models, including in the case of raw models (Zhao et al., 2024) where (most of) the task-level knowledge must be acquired during finetuning.

PEFT modules have been tested for cross-task transferability, with prompt tuning (Su et al., 2022; Vu et al., 2022), and other PEFT techniques, where e.g. Ding et al. (2023b) showed that PEFT-tuning, e.g. with LoRA, maintains performance on other tasks only when they are closely related.

Multiple LoRA modules have been used in mixture-of-expert set-ups, where the single most suitable module is selected on the fly for a given task (Feng et al., 2024; Dou et al., 2024). Recent work has also explored rank-wise clustering approaches to LoRA merging (Zhao et al., 2025). Most similar to our work, two papers have tested combining LoRA modules trained on general language tasks by computing the weighted sums of their parameters (with weights hardwired or optimised) (Asadi et al., 2024); or by (gradient-free) combinatorial optimisation for automatic selection of modules (Huang et al., 2023). The downsides are (i) additional learning steps after composition, and (ii) inefficient weight learning across many modules despite limited supervised data.

Our work differs from Asadi et al. (2024) in three key ways: (1) we introduce output composition methods which to our knowledge have not been previously explored for PEFT modules; (2) we focus on true plug-and-play composition without any post-composition training or optimization; and (3) we demonstrate that output composition not only generalises across related tasks but can even improve individual module performance on their original tasks.

The scenario we address is more ambitious: we want to take multiple task-trained LoRA modules off the shelf, combine them without further learning steps in a host model, to achieve performant results on each of the tasks for which we have loaded a module, as well as on combined tasks requiring their combined functionality. To the best of our knowledge, composability in this sense has not so far been explored; demonstrating it for QLoRA modules for the first time is our aim in this paper.

Our output composition techniques can be understood through the lens of representation engineering and activation steering (Turner et al., 2023; Zou et al., 2023), where model behaviour is controlled by manipulating internal representations. When we sum PEFT module outputs, we in effect combine learned steering directions in the model's representation space. Unlike typical activation steering approaches which require manual direction finding through techniques like contrastive activation addition or additional optimization steps (Li et al., 2023), our approach leverages task-specific PEFT modules as pre-learned steering vectors that can be composed without further training. This connection helps explain why output summing proves particularly effective: each PEFT module captures a task-specific representational direction, and their linear combination naturally integrates these learned behaviours.

3 Plug-and-play QLoRA

We reimplemented[1] QLoRA with additional functionality to support the attachment of multiple PEFT modules, and their composition with different techniques. In vanilla single-module QLoRA, a PEFT block consisting of a down-projection, and up-projection is attached (as shown in Figure 2 in Appendix) in parallel to every key, query, value and feed-forward layer in every transformer block.

Adopting the notation from Asadi et al. (2024), QLoRA learns a far smaller set of task-specific parameters $\Delta\Theta$ in conjunction with the frozen pre-trained model Θ_0. $\Delta\Theta$ (the QLoRA module) consists of trainable low-rank decomposition matrices added as adapter modules to the query, key, value and feed-forward weight matrices in the frozen model Θ_0. In other words, QLoRA adds task-specific parameters $\mathbf{\Delta W}$ to pre-trained frozen weight matrices $\mathbf{W}_0 \in \mathbb{R}^{d\times c}$ in Θ_0:

$$\hat{\mathbf{W}} = \mathbf{W_0} + \frac{\alpha}{r}\mathbf{\Delta W}$$

where $\mathbf{\Delta W} = \mathbf{AB^T}$ is the QLoRA block, with $\mathbf{A} \in \mathbb{R}^{d\times r}$ the down projection, $\mathbf{B}^T \in \mathbb{R}^{r\times d}$ the up projection, α a scaling factor, and r the rank.

$\mathbf{A}$ is initialised with the Kaiming uniform distribution, and $\mathbf{B}$ with all-zero initialisation. $\Delta\Theta$ is optimised with a standard conditional language modelling objective with CE loss computed over $\Theta_0 + \Delta\Theta$ (i.e. the frozen model with QLoRA blocks plugged in). For a given input x, the output h resulting from adapting the frozen parameters $\mathbf{W_0}$ with a QLoRA block is:

$$h = \mathbf{W_0}\, x + \frac{\alpha}{r}\mathbf{\Delta W} x$$

In our implementation of plug-and-play QLoRA, multiple blocks can be attached in each location and composed in one of three ways as detailed next.

[1] Available from https://github.com/anonymised

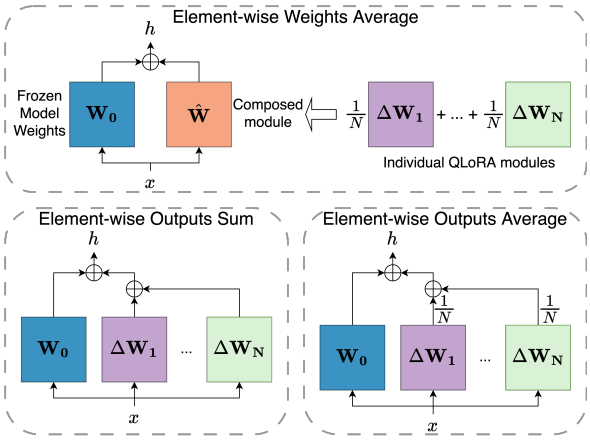

Figure 1: Diagram of our three QLoRA module composition techniques: (Top) Element-wise Weights Average; (Bottom-left) Element-wise Outputs Summing; (Bottom-right) Element-wise Outputs Averaging.

3.1 Composition techniques

Figure 1 shows the three composition techniques we tested in diagrammatic form. The first is **weights averaging**, where N PEFT modules are composed by computing the element-wise average of their weights (see top of Figure 1):

$$\hat{\mathbf{W}} = \frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{W_i}$$

where $\Delta\mathbf{W_i}$ are the weights of the i-th QLoRA module, and $\hat{\mathbf{W}}$ is the single module resulting from the composition. The output from the latter is summed to the output from the frozen parameters:

$$h = \mathbf{W_0}\,x + \frac{\alpha}{r}\hat{\mathbf{W}}x$$

The second composition technique is **output summing**, where the outputs of the N modules to be composed are first scaled by their respective scaling factors ($\frac{\alpha}{r} = \frac{16}{64}$ in all tests, see Section 6.2), and the element-wise sum of outputs (Figure 1, bottom left) is then computed and added to $\mathbf{W_0}\,x$:

$$h = \mathbf{W_0}\,x + \sum_{i=1}^{N}\frac{\alpha_i}{r_i}(\Delta\mathbf{W_i}\,x)$$

Output summing scales the total adapter contribution by the number of modules N. In our experiments we compose 2–5 modules; if tasks require substantially more modules to be composed, there may be effects on performance that we cannot predict from the experiments reported here.

Our third technique, **output averaging**, computes the element-wise average of module outputs (Figure 1, bottom right), before adding it to $\mathbf{W_0}x$:

$$h = \mathbf{W_0}\,x + \frac{1}{N}\sum_{i=1}^{N}\frac{\alpha_i}{r_i}(\Delta\mathbf{W_i}\,x)$$

Note that weight averaging and output averaging are not mathematically equivalent as further explained in Appendix C.

The theoretical advantage of output composition over weight composition lies in preserving learned module structure. Weight averaging merges the weight matrices themselves, losing the individual low-rank structures encoding task-specific information. Output summing computes each module's output before combining them, preserving task-specific representational directions that can be additively integrated in activation space, as shown by representation engineering findings (Section 2).

Our approach requires no additional training, but inference cost scales linearly with the number of modules N: output summing and averaging require N adapter forward passes for N modules. Weight averaging maintains the same proportional inference cost but has consistently weaker performance (Section 7), presenting a trade-off between computational efficiency and control effectiveness.

4 Study Overview

We aim to assess if the above composition techniques can achieve performant results on each of the individual module tasks, as well as on composite tasks requiring their combined functionality. We train QLoRA modules in conjunction with the frozen host model on individual tasks, compose them with one of the above three techniques, then test them on both individual and composite tasks.

For each pretrained raw model Θ, individual datasets $\mathbf{D} = \{D_1, ...D_m\}$, and composition techniques $\mathbf{C} = \{C_1, ...C_p\}$, we report results in terms of the schema shown in Table 1 which gives the results tables (Tables 2, 3, and 4) their structure and number of rows (#rows).

This study structure provides multiple points of comparison in terms of the performance of (i) the raw model on each dataset; (ii) the raw model + QLoRA module finetuned on each of the individual datasets, tested (a) on the data set it was finetuned on, and (b) on other related datasets; (iii) the raw model + QLoRA module finetuned on all individual datasets combined, tested on each separate data set; and (iv) the raw model + different types of composition applied to separately finetuned QLoRA modules, tested on the individual data sets.

QLoRA-tuned model (composition)	#rows
Raw model Θ	1
+ $\Delta\Theta_{D_i}$ trained on D_i	m
+ $\Delta\Theta_{\mathbf{D}}$ trained on combined $D_1 \cup ...D_m$	1
+ C_1 applied to subsets of $\Delta\Theta_{D_1}, ...\Delta\Theta_{D_m}$	$\sum_{k=2}^{m} \binom{m}{k}$
+ C_2 applied to subsets of $\Delta\Theta_{D_1}, ...\Delta\Theta_{D_m}$	$\sum_{k=2}^{m} \binom{m}{k}$
+ ...	
+ C_p applied to subsets of $\Delta\Theta_{D_1}, ...\Delta\Theta_{D_m}$	$\sum_{k=2}^{m} \binom{m}{k}$

Table 1: Study structure in terms of raw model and QLoRA (composed) modules. D_i = individual datasets; C_i = composition techniques. For other details, see text.

5 Models and Datasets

Models. We use three raw pretrained LLMs as host models: LLaMa 3 8B (AI@Meta, 2024), LLaMa 3.1 8B, and Mistral 7B (Jiang et al., 2023). For full details of models, see Appendix D.

Datasets. We use two sets of widely used single-task datasets, three for sentiment analysis which we use for sentiment-controlled text generation; and two for topic detection which we use for topic-controlled text generation. The ***sentiment*** datasets are Yelp Reviews (Chelba et al., 2013) (short review texts labelled pos/neg/neu), Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) (sentences extracted from reviews labelled with the sentence-level label pos/neg), and IMDb Movie Reviews (Maas et al., 2011) (longer review texts labelled pos/neg). The ***topic*** datasets are AG News (Zhang et al., 2015) (short news summaries labelled Sport, Business, Science/Tech, World); and DBpedia (Zhang et al., 2015) (very short entity descriptions with 14 entity class labels which we map to the AG News topics (as per Appendix B).

We use two short-text datasets for out-of-domain testing: PPLM Prompts (Dathathri et al., 2019), a collection of 35 2–4 word prompts, randomly chosen from sentence starters recommended for academic writing;[2] and the single-sentence image captions from the STS benchmark test set (Cer et al., 2017) from which we derive prompts in two ways: (i) using the whole sentence as a prompt (*STS* in tables), and (ii) using just the first n words as prompts, $n = 0..5$ (*STS proc* in tables). Neither of these datasets has attribute labels, so we generate and test one text for each control attribute label (combination) for each prompt in the dataset.

[2] http://www2.eit.ac.nz/library/ls_guides_sentencestarters.html

6 Experimental Set-up

6.1 Task construal

We use the above in-domain datasets for finetuning QLoRA modules on attribute-controlled prompted text generation. Prompts are leading text fragments tagged with the target attribute value(s), and the output is the text generated by the model in response, up to the first end-of-output tag, or a specified maximum length, whichever comes first.

Leading fragments are the first n words of a given text, where n ranges from 0 to 5 (2–4 for PPLM). Input-output pairs are then constructed by splitting the tagged text after n words, e.g. for sentiment-controlled generation control:

```
[SENTIMENT] Positive [\SENTIMENT] [ANS] The sushi is
great!  Not expensive & good quality.  My favorite rolles are the Vegas, Dragon, and the baked scallops :) [\ANS]
```

6.2 QLoRA module training

As a baseline, we finetune QLoRA modules on composite datasets constructed by filtering, balancing, and stratified sampling of multiple datasets from Section 5 to ensure size and label distribution consistency (see Appendix B for details).

A small grid search is performed for each model with the goal to balance affordable compute with model performance. Following the vanilla QLoRA hyperparameters (Dettmers et al., 2023), we set rank to 64, alpha to 16 and dropout to 0.1. QLoRA matrices are initialised using the Kaiming uniform distribution for A, using $a = \sqrt{5}$, and all-zero initialisation for B (see also Section 3). We train all our QLoRA modules in half-precision on an NVIDIA A100 with 80GB for 3 epochs. We save a checkpoint at each epoch, and select the one with the best Control Effectiveness (see Section 6.4) on the validation set. Each training/testing run is performed three times with different seeds.

We train one QLoRA module each for the five individual (Section 5) and two composite datasets (Section A).

6.3 Testing

We test the three composition techniques from Section 3.1 as per the study structure from Table 1, for two properties: (i) generalisation over related tasks, which here means all **sentiment control** tasks, or all **topic control** tasks; and (ii) functional composability, which here means the ability to control both sentiment and topic via module composition. The

CTG Technique	Distinct-n↑			SLOR↑	Control Effectiveness↑								
					Avg All	Avg	Yelp	IMDB	SST-2	Out-Of-Domain			
	dist-1	dist-2	dist-3							Avg	PPLM S	STS S	STS proc S
Llama 3 8B	0.03	0.09	0.12	9.44	59.93	63.70	68.00	57.56	65.56	57.16	64.60	52.83	54.06
+ QLoRA Yelp	0.09	0.24	0.34	9.87	85.11	91.44	92.78	90.67	90.89	79.49	88.41	**67.78**	82.28
+ QLoRA IMDB	0.04	0.13	0.19	**10.86**	83.07	92.07	92.11	89.67	94.44	73.61	80.95	62.00	77.89
+ QLoRA SST-2	**0.34**	**0.64**	**0.71**	8.24	77.80	85.96	84.11	82.44	91.33	70.73	82.86	57.39	71.94
+ QLoRA Combined Sentiment dataset	0.11	0.27	0.35	10.07	82.99	91.07	92.22	92.00	89.00	75.88	87.46	63.50	76.67
+ QLoRA Output Summing(IMDB, SST-2)	0.15	0.36	0.46	9.43	82.12	90.85	89.78	89.56	93.22	74.85	88.89	61.72	73.94
+ QLoRA Output Summing(IMDB, Yelp)	0.07	0.21	0.31	10.41	83.98	91.22	90.22	89.33	94.11	77.49	87.46	64.44	80.56
+ QLoRA Output Summing(Yelp, SST-2)	0.21	0.46	0.56	8.93	81.98	90.74	92.44	90.22	89.56	74.96	**90.00**	58.33	76.56
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	0.13	0.34	0.45	9.83	**85.66**	**92.48**	**94.44**	**92.56**	90.44	**79.54**	88.89	67.00	**82.72**
+ QLoRA Output Averaging(IMDB, SST-2)	0.12	0.27	0.35	9.49	79.29	88.30	86.33	87.44	91.11	71.24	83.65	58.44	71.61
+ QLoRA Output Averaging(IMDB, Yelp)	0.05	0.16	0.23	10.30	83.70	92.26	91.11	89.67	**96.00**	75.67	86.19	63.17	77.67
+ QLoRA Output Averaging(Yelp, SST-2)	0.13	0.31	0.38	9.16	81.96	89.74	89.00	88.56	91.67	74.69	84.29	62.61	77.17
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	0.07	0.19	0.27	9.77	81.04	89.81	88.78	89.00	91.67	73.03	84.60	60.50	74.00
+ QLoRA Averaged Weights(IMDB, SST-2)	0.31	0.61	0.69	8.37	77.50	85.93	85.33	81.89	90.56	69.16	77.94	57.50	72.06
+ QLoRA Averaged Weights(IMDB, Yelp)	0.04	0.12	0.19	10.83	81.96	90.74	91.89	87.22	93.11	73.65	84.13	60.67	76.17
+ QLoRA Averaged Weights(Yelp, SST-2)	0.33	**0.64**	**0.71**	8.28	77.94	85.93	85.33	81.89	90.56	71.44	84.76	57.50	72.06
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	0.11	0.28	0.36	9.54	80.29	90.04	88.89	88.78	92.44	71.17	83.17	58.72	71.61
Llama 3.1 8B	0.04	0.10	0.13	9.66	58.32	61.52	66.22	54.22	64.11	56.41	64.29	51.56	53.39
+ QLoRA Yelp	0.09	0.26	0.36	9.94	86.68	91.52	91.00	89.00	94.56	**82.93**	91.75	**71.28**	**85.78**
+ QLoRA IMDB	0.04	0.14	0.21	**10.81**	83.46	92.26	89.67	91.44	**95.67**	75.04	85.24	63.33	76.56
+ QLoRA SST-2	**0.36**	**0.69**	**0.76**	8.24	78.29	85.48	82.33	82.89	91.22	72.49	84.76	58.17	74.56
+ QLoRA Combined Sentiment dataset	0.12	0.27	0.35	10.00	84.75	92.67	94.89	91.44	91.67	77.86	89.52	62.28	81.78
+ QLoRA Output Summing(IMDB, SST-2)	0.09	0.23	0.31	9.91	84.98	92.93	93.44	90.89	94.44	78.68	**92.54**	62.78	80.72
+ QLoRA Output Summing(IMDB, Yelp)	0.07	0.21	0.31	10.33	85.50	91.52	91.22	90.00	93.33	80.26	89.05	68.06	83.67
+ QLoRA Output Summing(Yelp, SST-2)	0.16	0.38	0.47	9.16	85.38	92.48	92.78	91.44	93.22	79.07	89.05	68.94	79.22
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	0.10	0.27	0.37	9.91	**87.35**	**94.70**	**96.33**	**92.67**	95.11	80.61	90.16	66.89	84.78
+ QLoRA Output Averaging(IMDB, SST-2)	0.08	0.21	0.28	9.81	81.14	89.74	89.22	87.22	92.78	72.95	83.02	61.50	74.33
+ QLoRA Output Averaging(IMDB, Yelp)	0.07	0.22	0.31	10.24	83.45	92.30	93.00	90.22	93.67	75.16	86.03	63.22	76.22
+ QLoRA Output Averaging(Yelp, SST-2)	0.16	0.36	0.44	9.05	80.54	89.00	89.67	86.89	90.44	73.06	85.08	60.94	73.17
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	0.07	0.18	0.26	10.01	79.79	89.15	88.56	87.11	91.78	71.69	85.56	58.39	71.11
+ QLoRA Averaged Weights(IMDB, SST-2)	0.33	0.64	0.71	8.38	76.87	84.07	81.00	81.89	89.33	71.21	83.97	56.67	73.00
+ QLoRA Averaged Weights(IMDB, Yelp)	0.04	0.14	0.21	10.78	82.24	91.19	90.67	88.89	94.00	73.57	83.49	61.89	75.33
+ QLoRA Averaged Weights(Yelp, SST-2)	0.35	0.67	0.74	8.27	76.83	84.07	81.00	81.89	89.33	71.00	83.33	56.67	73.00
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	0.08	0.20	0.27	9.87	81.48	89.67	91.67	86.67	90.67	74.19	85.56	60.72	76.28
Mistral 7B	0.04	0.09	0.12	7.45	57.54	60.00	59.33	55.00	65.67	56.23	62.86	52.50	53.33
+ QLoRA Yelp	0.03	0.07	0.10	10.44	**84.33**	92.30	**94.78**	**90.56**	91.56	**76.42**	84.60	**61.33**	**83.33**
+ QLoRA IMDB	0.02	0.04	0.06	**11.25**	80.24	90.63	89.67	88.56	93.67	70.38	82.70	58.78	69.67
+ QLoRA SST-2	**0.27**	**0.51**	**0.58**	7.65	78.17	85.67	83.11	81.67	92.22	71.87	83.65	55.94	76.00
+ QLoRA Combined Sentiment dataset	0.03	0.06	0.09	10.83	78.28	86.30	87.11	85.67	86.11	71.32	83.17	58.22	72.56
+ QLoRA Output Summing(IMDB, SST-2)	0.22	0.41	0.47	7.80	79.28	87.11	85.56	85.44	90.33	73.12	86.98	57.11	75.28
+ QLoRA Output Summing(IMDB, Yelp)	0.03	0.07	0.10	10.51	82.53	90.44	93.33	86.33	91.67	75.59	86.98	58.28	81.50
+ QLoRA Output Summing(Yelp, SST-2)	0.18	0.35	0.40	8.22	80.17	88.26	89.00	82.33	93.44	74.13	**89.68**	59.22	73.50
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	0.08	0.16	0.19	9.13	83.06	**92.81**	93.78	89.44	**95.22**	74.09	86.67	58.89	76.72
+ QLoRA Output Averaging(IMDB, SST-2)	0.11	0.21	0.25	8.49	79.09	88.89	88.56	85.33	92.78	69.93	82.06	58.39	69.33
+ QLoRA Output Averaging(IMDB, Yelp)	0.02	0.06	0.09	10.80	82.23	92.63	93.56	90.11	94.22	72.35	84.60	56.78	75.67
+ QLoRA Output Averaging(Yelp, SST-2)	0.05	0.10	0.13	8.84	80.80	90.74	92.44	86.33	93.44	71.96	85.87	60.67	69.33
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	0.02	0.05	0.07	10.54	81.31	92.37	91.67	**90.56**	94.89	70.99	84.76	57.67	70.56
+ QLoRA Averaged Weights(IMDB, SST-2)	0.25	0.49	0.55	7.69	78.22	85.85	82.89	82.22	92.44	71.98	84.60	56.28	75.06
+ QLoRA Averaged Weights(IMDB, Yelp)	0.02	0.04	0.07	11.19	80.71	91.33	90.22	89.89	93.89	70.72	83.65	59.56	68.94
+ QLoRA Averaged Weights(Yelp, SST-2)	**0.27**	0.50	0.57	7.66	78.35	85.85	82.89	82.22	92.44	72.61	86.51	56.28	75.06
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	0.10	0.20	0.24	8.64	79.19	88.26	88.11	84.11	92.56	71.67	86.35	58.00	70.67

Table 2: **Sentiment Control** Diversity, Fluency, Control Effectiveness for the model + QLoRA module combinations explained in Table 1. Here, e.g. Output Summing(data1, data2) refers to the output summation module composition technique. All values are averages over 3 runs. Standard deviations are reported in Appendix Tables 11 and 12. Bold (shaded) = (second) highest score in column/section; underline = train/test on same dataset.

latter is a **multiple-attribute control** task, and here we test two distinct composition strategies: (i) composing the two single-attribute modules trained on the composite Combined Sentiment/Topic datasets; and (ii) composing the five single-attribute modules trained on the separate individual datasets.

6.4 Evaluation metrics

We measure ***Diversity*** with Distinct-n (Li et al., 2015) which is the proportion of unique n-grams in generated texts. The aggregate Distinct-n score is the mean of item-level scores.

We assess ***Fluency*** with average Syntactic Log-Odds Ratio (SLOR) (Kann et al., 2018) obtained with GPT-2XL and BLOOM 1B7, in preference over perplexity which may not effectively capture fluency for low-frequency items. SLOR calculates the log-probability of a sentence normalised by unigram log-probability and length.

To assess ***Control Effectiveness*** (CE), we mea-

CTG Technique	Distinct-n↑			SLOR↑	Control Effectiveness↑							
					Avg All	Avg	AG News	DBPedia	Out-Of-Domain			
	dist-1	dist-2	dist-3						Avg	PPLM T	STS T	STS proc T
Llama 3 8B	0.07	0.17	0.22	8.45	45.47	58.52	64.61	52.42	37.53	48.97	27.97	35.64
+ QLoRA AG News	0.25	0.53	0.62	9.39	**68.53**	**85.13**	**90.72**	**79.54**	**58.02**	**71.11**	**42.03**	60.92
+ QLoRA DBPedia	0.32	0.60	0.68	8.96	52.61	69.94	71.67	68.21	41.66	55.40	28.58	41.00
+ QLoRA Combined Topic dataset	0.30	0.60	0.70	8.92	62.76	74.42	81.61	67.24	56.27	68.73	37.50	**62.58**
+ QLoRA Output Summing(AG News, DBPedia)	**0.35**	**0.70**	**0.80**	9.33	63.61	82.57	88.78	76.35	52.59	**71.11**	32.81	53.86
+ QLoRA Output Averaging(AG News, DBPedia)	0.27	0.55	0.64	**9.57**	60.97	78.33	83.78	72.88	50.70	66.98	33.81	51.31
+ QLoRA Averaged Weights(AG News, DBPedia)	0.33	0.62	0.71	8.98	53.90	70.61	72.11	69.12	45.50	66.59	28.06	41.86
Llama 3.1 8B	0.05	0.12	0.17	**9.45**	45.86	58.61	66.11	51.11	37.47	46.35	29.89	36.17
+ QLoRA AG News	0.26	0.55	0.65	9.40	**67.85**	**85.52**	**91.33**	**79.72**	**57.21**	**73.10**	36.28	**62.25**
+ QLoRA DBPedia	0.31	0.59	0.68	8.74	52.63	69.86	70.94	68.77	41.64	54.92	28.11	41.89
+ QLoRA Combined Topic dataset	0.31	0.62	0.73	9.29	65.08	80.65	89.06	72.25	55.01	66.51	**39.97**	58.56
+ QLoRA Output Summing(AG News, DBPedia)	**0.35**	**0.68**	**0.78**	9.12	63.41	82.91	88.39	77.44	51.64	69.05	29.39	56.47
+ QLoRA Output Averaging(AG News, DBPedia)	0.30	0.59	0.68	8.98	58.85	76.42	83.44	69.40	47.74	61.67	29.81	51.75
+ QLoRA Averaged Weights(AG News, DBPedia)	0.32	0.61	0.71	8.74	52.59	69.52	70.94	68.09	43.06	60.71	27.61	40.86
Mistral 7B	0.07	0.13	0.17	7.98	44.14	55.96	62.00	49.91	37.48	49.76	29.92	32.75
+ QLoRA AG News	0.18	0.37	0.44	9.48	**69.05**	**88.52**	**93.17**	**83.87**	**57.14**	**73.97**	**39.69**	**57.75**
+ QLoRA DBPedia	**0.26**	0.48	0.56	8.81	52.92	67.05	67.50	66.61	43.76	54.13	32.89	44.28
+ QLoRA Combined Topic dataset	0.21	0.43	0.52	9.09	64.64	81.74	89.00	74.47	53.87	67.54	39.22	54.86
+ QLoRA Output Summing(AG News, DBPedia)	**0.26**	**0.50**	**0.59**	9.26	65.46	85.27	88.89	81.65	53.19	69.76	36.36	53.44
+ QLoRA Output Averaging(AG News, DBPedia)	0.18	0.36	0.42	**9.60**	58.01	76.01	82.00	70.03	46.64	60.95	33.89	45.08
+ QLoRA Averaged Weights(AG News, DBPedia)	**0.26**	0.48	0.56	8.87	53.58	66.53	66.06	67.01	46.72	61.75	33.47	44.94

Table 3: Diversity, Fluency, Control Effectiveness for **Topic Control**, training on *single* and *combined* datasets, and composition of modules trained on single datasets, e.g. Output Summing(data1, data2). All values are averages over 3 runs. Standard deviations are reported in Appendix Tables 13 and 14. Bold (shaded) = (second) highest score in column and section; underline = train and test set from same dataset.

sure the mean percentage of texts identified by a set of classifiers as possessing the controlled attribute value. I.e. we first compute the percentage of cases for each classifier where the detected attribute value matches the input control value; the final CE score is then the average of the three classifiers' individual percentages. We use DistilBERT and T5 fine-tuned on SST-2, and DeBERTa fine-tuned on Yelp, for sentiment control; and DistilBERT, BERT, and DeBERTa fine-tuned on AG-News, for topic control (Appendix H.2). In multiple-attribute control, instead of one target attribute value having to be matched, both have to be matched. We use majority voting across classifiers to obtain sentiment and topic labels before calculating CE as above.

7 Results

7.1 Sentiment control

Table 2 presents results for sentiment control in terms of the row structure introduced in Table 1. Across the columns, we have results for Diversity measured by Distinct-n, Fluency by SLOR, and Control Effectiveness by classifier average (see Section 6.4), on in-domain and out-of-domain datasets (Section 5), as well as averaged over each.

Regarding *Diversity*, strikingly, training on SST-2 always achieves the most diverse outputs by a very substantial margin, across all three models, although Weights Averaging closely matches this if the SST-2 trained module is in the composition. For *Fluency*, it is training on IMDb that achieves the highest scores, albeit by smaller margins.

For *Control Effectiveness*, while the three (unchanged) raw models perform on average on a par (first row in each section), Mistral responds slightly worse to QLoRA-tuning across all settings. We observe a clear trend where Output Summing consistently achieves the highest scores for module compositions, except for Mistral, where results present a slightly more mixed picture. Despite this, the Summing technique achieves best or second-best results in out-of-domain testing across all models.

Regarding individual train/test set combinations, strikingly it is rarely the case that the model trained on a single dataset achieves the best result on it. For the Llama models, when training on the combined data, performance is improved or maintained compared to training on the same data set in all cases but one (SST-2). In contrast, for Mistral, performance always drops considerably.

For the last three columns (out-of-domain testing), the Yelp-trained model versions always achieve the best result, in most cases by considerable margins, with the overall best results achieved by Llama 3.1 8B + QLoRA Yelp.

These results present a clear overall trend where our new output composition approaches (Summing and Averaging) consistently outperform the Weights Average method across all models and

CTG Technique	Distinct-n↑			SLOR↑	Control Effectiveness ↑									
					Out-Of-Domain									
					Multiple				Sentiment			Topic		
	dist1	dist2	dist3		Avg	PPLM M	STS M	STS p M	PPLM S	STS S	STS p S	PPLM T	STS T	STS p T
Llama 3 8B	0.04	0.11	0.16	9.77	19.22	29.29	14.54	20.38	64.60	52.83	54.06	48.97	27.97	35.64
+ QLoRA Combined Sentiment dataset	0.03	0.09	0.14	**10.81**	**25.66**	37.14	16.50	**30.79**	87.46	**63.50**	**76.67**	51.67	27.69	38.36
+ QLoRA Combined Topic dataset	**0.24**	**0.50**	**0.59**	8.52	22.64	36.55	14.96	25.46	55.71	50.78	54.83	**68.73**	**37.50**	**62.58**
+ QLoRA Output Summing(S, T)	0.09	0.23	0.31	9.81	23.85	35.60	**17.75**	25.83	**87.62**	59.56	72.72	62.46	29.42	50.25
+ QLoRA Output Summing(Ind mod)	0.18	0.42	0.51	9.11	24.08	33.69	16.79	28.00	79.68	59.22	69.39	43.97	25.22	37.75
+ QLoRA Output Averaging(S, T)	0.10	0.24	0.31	9.26	22.32	36.55	14.50	25.17	76.35	54.56	63.83	57.86	29.39	42.11
+ QLoRA Output Averaging(Ind mod)	0.08	0.20	0.26	9.55	23.30	**37.50**	15.46	26.17	78.41	57.17	67.61	57.94	26.58	43.22
+ QLoRA Averaged Weights(S, T)	0.21	0.46	0.55	8.56	22.85	36.55	15.62	25.29	74.92	50.78	54.83	67.70	36.83	59.86
+ QLoRA Averaged Weights(Ind mod)	0.09	0.23	0.30	9.45	23.58	35.71	15.79	27.12	73.65	54.50	63.17	64.84	32.11	46.33
Llama 3.1 8B	0.05	0.12	0.18	9.84	19.95	28.57	15.00	21.88	64.29	51.56	53.39	46.35	29.89	36.17
+ QLoRA Combined Sentiment dataset	0.03	0.11	0.17	**10.73**	26.44	37.50	17.54	**31.46**	**89.52**	62.28	**81.78**	49.13	30.72	36.86
+ QLoRA Combined Topic dataset	**0.28**	**0.58**	**0.68**	8.58	23.95	35.12	17.54	26.46	52.06	51.39	51.83	66.51	**39.97**	58.56
+ QLoRA Output Summing(S, T)	0.10	0.27	0.36	9.80	**26.76**	34.40	**19.58**	31.25	75.56	**62.89**	71.83	58.17	34.97	49.75
+ QLoRA Output Summing(Ind mod)	0.13	0.29	0.35	8.63	20.78	30.24	16.29	21.96	84.60	60.94	72.94	35.00	24.42	28.92
+ QLoRA Output Averaging(S, T)	0.11	0.29	0.37	9.67	25.89	**40.83**	16.50	30.04	73.81	59.83	63.72	59.44	29.61	43.08
+ QLoRA Output Averaging(Ind mod)	0.08	0.22	0.29	9.36	23.17	37.38	15.21	26.17	69.84	54.50	64.89	45.87	27.03	39.42
+ QLoRA Averaged Weights(S, T)	0.24	0.52	0.62	8.74	24.43	37.98	17.29	26.83	73.49	51.39	51.83	**70.08**	39.61	**61.11**
+ QLoRA Averaged Weights(Ind mod)	0.10	0.25	0.33	9.14	23.26	40.36	15.46	25.08	72.22	53.61	64.06	50.40	28.28	39.81
Mistral 7B	0.03	0.07	0.10	9.77	20.43	26.07	16.12	22.75	62.86	52.50	53.33	49.76	29.92	32.75
+ QLoRA Combined Sentiment dataset	0.01	0.03	0.05	**11.60**	21.47	30.60	14.75	25.00	83.17	**58.22**	72.56	46.83	27.44	38.33
+ QLoRA Combined Topic dataset	**0.12**	**0.27**	**0.36**	8.92	26.06	35.12	20.33	28.62	50.16	51.50	52.33	67.54	39.22	54.86
+ QLoRA Output Summing(S, T)	0.09	0.22	0.30	9.50	25.53	34.05	19.58	28.50	59.84	51.33	57.56	63.33	39.75	**60.88**
+ QLoRA Output Summing(Ind mod)	0.09	0.20	0.24	8.94	22.77	36.19	16.25	24.58	**87.78**	57.33	**75.83**	57.30	31.44	42.14
+ QLoRA Output Averaging(S, T)	0.08	0.18	0.23	9.63	22.73	34.88	15.54	25.67	63.17	52.83	57.17	60.87	29.92	43.97
+ QLoRA Output Averaging(Ind mod)	0.02	0.05	0.08	10.75	23.01	35.95	14.08	27.42	76.19	54.78	66.89	48.17	27.81	37.61
+ QLoRA Averaged Weights(S, T)	0.10	0.24	0.32	9.04	**26.24**	27.14	**21.25**	**30.92**	59.05	51.50	52.33	**70.00**	**39.92**	55.78
+ QLoRA Averaged Weights(Ind mod)	0.03	0.07	0.10	10.18	21.70	**36.31**	14.42	23.88	77.46	54.28	62.44	52.86	28.00	38.78

Table 4: Diversity, Fluency, Control Effectiveness for **Multi-attribute Control** alongside single-attribute control results for comparison. All values are averages over 3 runs. Standard deviations are reported in Appendix Tables 15 and 16. S=the Combined Sentiment dataset, T=the Combined Topic dataset, M=Multiple, p = processed, Ind mod=composition is on all 5 individually trained modules. Bold (shaded) = (second) highest score in column and section; underline = train and test set from same dataset.

evaluation settings, indicating that combining the outputs of separately fine-tuned PEFT modules is a more effective strategy for achieving control effectiveness than averaging their weights.

7.2 Topic control

Table 3 presents results for topic control using the same structure and notation as in Table 2. *Diversity* scores here are generally higher than for sentiment. Summing the two modules performs best for all settings. *Fluency* scores are very similar across the different settings with no notable trends emerging.

In *Control Effectiveness*, for both in-domain and out-of-domain datasets, the single AG News trained module (second row in each section) performs best in all but two cases. The summed modules have the second highest scores in 11 cases, and the single module trained on the Combined dataset has the second highest scores also in 11 cases.

Among the composition techniques, we confirm the consistent trend where Output Summing outperforms the other methods.

Note that training on AG News has a twofold advantage over DBPedia in the current context: (i) the classifiers in the CE metric were all trained on AG News, and (ii) the mapping from topics in DBPedia to AG News labels is imperfect resulting in a noisier text-label relationship. To assess the impact we conducted an additional evaluation using an instruction-tuned LLM as a classifier which confirmed the findings in Table 3, with both Pearson's and Spearman's correlations around or above 0.9 (see Appendix E for details). This suggests no adverse effect from classifier bias.

7.3 Multiple-attribute control

The single-attribute results so far presented shed light on the ability of module composition to generalise across individual datasets. The results for multiple-attribute control over both sentiment and topic presented in this section test the functional composability of the modules.

Table 4 presents the multiple-attribute control results. We can see that in terms of *Diversity*, modules trained on (just) the Combined Topic dataset achieve the highest results for all models. In terms of *Fluency*, all models perform best with the module trained on the Combined Sentiment dataset.

For *Control Effectiveness* overall, finetuning consistently performs above raw model baseline for the Llama models, by big margins. For Mistral, finetuned models in some cases perform worse.

For Multiple-control, results present a mixed picture. One perspective is provided by comparing modules trained on the Combined datasets with those composing modules trained on individual datasets. Here, results are dataset and model dependent: for Llama, Output Summing is best for STS M, Output Averaging is best for PPLM M, and Combined dataset training is best for STS p M; for Mistral generally, Weights Averaging is best.

The finetuned models in this section were all task-specifically created for multiple-attribute control, so the above are the main results in this section. However, we also wanted to see to what extent the multi-attribute control models retain their single-attribute control performance, shown in the last six columns in Table 4. A clear trend is that best results are worse than in single-attribute control (Tables 2 and 3). Which composition methods work best also follows clear trends: for retaining sentiment control performance, it is the module trained on the Combined Topic dataset combined with Weights Averaging. For retaining topic control performance, it is the module trained on the Combined Sentiment dataset combined with Output Summing.

For retention of topic control performance, Weights Averaging yields best or second-best results across all models. However, with Mistral, our new Summing approach achieves comparable or higher performance than Weights Averaging, suggesting that output-based composition is also competitive in this scenario under certain conditions.

8 Discussion

State-of-the-art controlled-generation methods like Prior CTG (Gu et al., 2023), PPLM (Dathathri et al., 2019), and GeDi (Krause et al., 2021) report control effectiveness in the range of 80–97% and 74–97% for sentiment and topic control, compared to our ranges of 86–96% and 70–93%, respectively. For multi-attribute control, average single-attribute performance tends to be reported, with results typically in the range 71–92%. This is clearly not comparable to our results where we use three classifiers and require all predicted labels to be correct. More comparably, for MacLaSa, Ding et al. (2023a) compute CE as the percentage of times both predicted labels are correct, reporting considerably lower effectiveness, ranging from 18–59%, better reflecting the difficulty of controlling multiple attributes.

Across all experiments, Output Summing consistently outperforms other composition techniques. Our results indicate that summing the output of multiple PEFT modules effectively preserves the knowledge captured in individual modules, particularly when composing modules trained on the same control attribute. The success of summing (and to a lesser extent, averaging) module outputs suggests that trained modules project inputs into the same latent space effectively.

When combining modules trained on different tasks or control attributes, we observe that composed modules successfully preserve single-attribute knowledge, particularly for sentiment. However, topic control is less effective, suggesting that the signal from topic modules may be less strong, leading to weaker control over diverse attributes. A possible improvement could be weighted module composition, where specific attributes are reinforced to ensure stronger control. Despite this, our findings confirm that even when composing modules that have been fine-tuned for different single-attribute tasks, the model retains its single-attribute control capabilities.

Three-module output summing outperforms task-specialized single modules in 7 of 9 scenarios, demonstrating that composition not only preserves but enhances single-task performance through cross-dataset generalization (see Appendix F for detailed analysis).

9 Conclusion

We have presented a first-time examination of the output composability of multiple QLoRA modules within the same host model, in a fully plug-and-play setting. We set out to assess two tests of composability: (i) generalisation over related tasks, and (ii) functional composability on composite tasks. Re *i*, our new output summing composition method consistently provided the best performance, not only generalising well over multiple tasks, but even improving performance on the individual tasks, averaging a 2% *improvement* over single-task trained modules on the corresponding single task test set. Re *ii*, module composition overall provided an advantage, but could not always outperform training a single module on data set combinations. Our results provide a first demonstration of the astonishing composability of QLoRA-finetuned modules.

Limitations

In this work, we focus on exploring our approach only on three pre-trained raw models with similar number of parameters, same number of layers and same architecture type. While this choice allows us to isolate and test the effects of our approach, it restricts our understanding of its general applicability. Testing our approach on models with different number of parameters, architecture type, and number of layers would give a more comprehensive picture of its robustness and versatility.

Our evaluation is limited to two control attributes, sentiment and topic, across related datasets within the same domain. While this focused scope allows us to systematically isolate composition mechanics, it limits the generality of our conclusions. Whether the composability patterns we observe, particularly the effectiveness of output summing, extend to (i) more complex generative behaviours such as reasoning or creative writing, (ii) fundamentally different task types (e.g., combining classification, question-answering, and generation modules), or (iii) unrelated domains remains unexplored. These represent important directions for future work to establish the broader applicability of PEFT module composition.

Our experiments compose 2-5 modules trained on related tasks within the same domain. Real-world applications may require orchestrating dozens or hundreds of modules across truly disparate tasks (e.g., combining modules for translation, summarization, code generation, and question-answering). Such large-scale scenarios introduce several challenges we do not address: (1) computational cost scaling linearly with active modules (O(N)) could become prohibitive; (2) output summing scales adapter contributions linearly with N, potentially causing output magnitudes to deviate significantly from training distributions; (3) potential interference effects between unrelated task modules; and (4) determining which modules to activate for complex queries. Addressing large-scale heterogeneous composition would likely require additional mechanisms such as: dynamic module selection or mixture-of-experts-style routing to activate only relevant subsets; normalization or learned weighting schemes to modulate individual contributions and control output magnitude; adaptive scaling factors; or meta-learning to predict module compatibility and optimal composition strategies.

Additionally, our stratified sampling approach ensures all modules are trained on equal-sized datasets, which enables fair comparison of composition techniques but does not explore how output summing performs with varying training data sizes, an important consideration for real-world scenarios where modules may be trained on datasets of vastly different scales.

Furthermore, we limit this work to investigating only one PEFT technique, which leaves open the question of whether our approach would perform the same with other PEFT techniques. Expanding the scope of our work to include different PEFT techniques could help evaluate the generality of the approach. Furthermore, incorporating the composability of modules generated by different PEFT techniques would shed light on the extent to which our method is compatible with mixed-module designs.

Finally, our evaluation setup relies on automatic metrics to assess the model's performance. A dedicated human evaluation would provide a more nuanced understanding of the generated texts.

Ethical Considerations

We test our approach on two well-known control attributes, namely sentiment and topic. However, our approach could potentially be used to force the model to include and generate offensive, inappropriate, or biased content. At present, our approach does not include any built-in mechanisms to prevent or mitigate such misuse, highlighting an important area for future work in ensuring ethical and responsible application.

Additionally, our approach is based on the use of pre-trained LLMs, which are known to inherit biases and limitations from the data they were trained on. The generated outputs might include offensive, incorrect, or harmful content.

References

AI@Meta. 2024. Llama 3 model card.

Nader Asadi, Mahdi Beitollahi, Yasser H Khalil, Yinchuan Li, Guojun Zhang, and Xi Chen. 2024. Combining pre-trained lora modules improves few-shot adaptation of foundation models to new tasks. In *ICML 2024 Workshop on Foundation Models in the Wild*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings*

of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. In *arXiv preprint arXiv:1312.3005*.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023a. MacLaSa: Multi-aspect controllable text generation via efficient sampling from compact latent space. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4424–4436, Singapore. Association for Computational Linguistics.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023b. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3):220–235.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Wei Shen, Limao Xiong, Yuhao Zhou, Xiao Wang, Zhiheng Xi, Xiaoran Fan, Shiliang Pu, Jiang Zhu, Rui Zheng, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. LoRAMoE: Alleviating world knowledge forgetting in large language models via MoE-style plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1932–1945, Bangkok, Thailand. Association for Computational Linguistics.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Yu Han, and Hao Wang. 2024. Mixture-of-LoRAs: An efficient multitask tuning method for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11371–11380, Torino, Italia. ELRA and ICCL.

Yuxuan Gu, Xiaocheng Feng, Sicheng Ma, Lingyuan Zhang, Heng Gong, Weihong Zhong, and Bing Qin. 2023. Controllable text generation via probability density estimation in the latent space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12590–12616, Toronto, Canada. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Chengsong Huang, Qian Liu, Bill Yuchen Lin, Tianyu Pang, Chao Du, and Min Lin. 2023. Lorahub: Efficient cross-task generalization via dynamic lora composition. *arXiv preprint arXiv:2307.13269*.

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

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2021. GeDi: Generative discriminator guided sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929–4952, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2024. Gpt understands, too. *AI Open*, 5:208–215.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Clifton Poth, Hannah Sterz, Indraneil Paul, Sukannya Purkayastha, Leon Engländer, Timo Imhof, Ivan Vulić, Sebastian Ruder, Iryna Gurevych, and Jonas Pfeiffer. 2023. Adapters: A unified library for parameter-efficient and modular transfer learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 149–160.

Mohammed Sabry and Anya Belz. 2024. Assessing the portability of parameter matrices trained by parameter-efficient finetuning methods. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1548–1556, St. Julian's, Malta. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew 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*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, Lei Hou, Maosong Sun, and Jie Zhou. 2022. On transferability of prompt tuning for natural language processing. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3949–3969, Seattle, United States. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *CoRR*.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou', and Daniel Cer. 2022. SPoT: Better frozen model adaptation through soft prompt transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5039–5059, Dublin, Ireland. Association for Computational Linguistics.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1202–1228, Mexico City, Mexico. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Justin Zhao, Timothy Wang, Wael Abid, Geoffrey Angus, Arnav Garg, Jeffery Kinnison, Alex Sherstinsky, Piero Molino, Travis Addair, and Devvret Rishi. 2024. Lora land: 310 fine-tuned llms that rival gpt-4, a technical report. *arXiv preprint arXiv:2405.00732*.

Ziyu Zhao, Tao Shen, Didi Zhu, Zexi Li, Jing Su, Xuwu Wang, and Fei Wu. 2025. Merging loRAs like playing LEGO: Pushing the modularity of loRA to extremes through rank-wise clustering. In *The Thirteenth International Conference on Learning Representations*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *CoRR*.

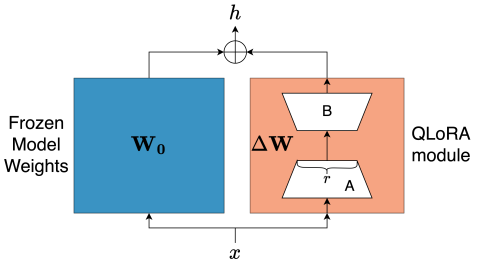

Figure 2: A single QLoRA block (orange) shown attached to its corresponding model weights.

A Composite dataset construction

As a baseline point of comparison we finetune QLoRA modules on a composite dataset made up of multiple (subsets of) datasets from Section 5, created via the following three-step process:

1. **Filtering**: Dataset items with fewer than 10 words, and any neutral-tagged sentiment dataset items (just Yelp) are removed.

2. **Smallest dataset balancing**: We identify the smallest dataset; if its label distribution is imbalanced, it is randomly downsampled to get a balanced distribution of items over labels.

3. **Stratified sampling**: All other datasets are then downsampled to match the size of the smallest dataset, ensuring that each dataset is also balanced across labels and text length.

Dataset	Split	Before sampling			After sampling		
		Num samples	Avg words	Avg char	Num samples	Avg words	Avg char
IMDb	train	25000	231.49	1302.62	24330	231.50	1302.54
	test	25000	226.25	1243.82	100	218.61	1271.64
SST-2	train	67349	9.41	97.21	24330	17.86	53.03
	validation	872	19.55	116.24	726	21.62	105.05
	test	1821	19.23	49.86	100	9.33	103.15
Yelp	train	470000	134.03	728.80	24330	135.85	718.88
	validation	50000	133.87	746.85	726	138.81	717.61
	test	40000	133.87	712.02	100	132.03	718.08
Combined Sentiment	train	-	-	-	72990	128.41	709.54
	validation	-	-	-	1452	80.22	431.55
AG News	train	112000	37.85	236.11	111992	37.86	236.10
	validation	8000	37.78	235.75	3996	37.78	235.67
	test	7600	37.72	231.37	200	37.42	234.90
DBpedia	train	546000	46.13	275.05	111984	44.91	280.56
	validation	14000	46.16	274.94	3996	44.90	280.49
	test	70000	46.14	265.66	195	43.12	280.57
Combined Topic	train	-	-	-	223976	41.38	255.58
	validation	-	-	-	7992	41.34	255.35
PPLM	test	35	2.54	13.31	35	2.54	13.31
STS	test	625	8.02	39.36	100	7.50	36.83
STS proc	test	-	-	-	100	3.18	13.71

Table 5: Statistics of the datasets used for each data split (train, validation, test), showing the number of samples, average number of words per text, and average number of characters per text, both before and after dataset sampling.

Original label	Mapped label
Company	Business
Educational	World
Artist	World
Athlete	Sports
OfficeHolder	Business
MeanOfTransportation	World
Building	World
NaturalPlace	World
Village	World
Animal	Sci/Tech
Plant	Sci/Tech
Album	World
Film	World
WrittenWork	World

Table 6: Mapping of DBPedia topic labels into AG News topic labels

We combine the three sampled sentiment datasets to form the Combined Sentiment dataset, and the two topic datasets to form the Combined Topic dataset. For these combined datasets, we sample only from the training and validation splits. In contrast, for the single-dataset test sets, we include up to 50 examples per label. For descriptive statistics of datasets see Appendix B.

B Dataset Descriptives

Table 5 shows a detailed overview of the datasets used, including details of each data split (train, validation, and test). The table reports the number of samples, the average number of words per text, and the average number of characters per text before and after stratified sampling. We can notice that the stratified sampling based on text length effectively created balanced datasets while preserving the diversity of input length.

As discussed in Section 5, it is necessary to align the topic labels in the DBPedia dataset with the topic labels in the AG News dataset to ensure a consistent set of topics across both datasets. Table 6 shows the mapping from the original DBPedia topics to the AG News topics.

CTG Technique	Control Effectiveness↑							
	Avg All	Avg	AG News	DBPedia	Out-Of-Domain			
					Avg	PPLM T	STS T	STS proc T
Llama 3 8B	54.49	73.22	75.67	70.77	43.25	60.24	27.92	41.58
+ QLoRA AG News	**68.82**	**84.56**	**84.50**	**84.62**	**59.72**	**75.24**	**42.67**	61.25
+ QLoRA DBPedia	59.43	78.31	78.83	77.78	48.21	65.71	31.17	47.75
+ QLoRA Combined Topic dataset	63.51	75.19	75.50	74.87	58.22	75.00	37.00	**62.67**
+ QLoRA Output Summing(AG News, DBPedia)	64.44	83.02	84.33	81.71	53.87	72.86	35.17	53.58
+ QLoRA Output Averaging(AG News, DBPedia)	63.05	81.60	81.67	81.54	52.76	72.62	32.83	52.83
+ QLoRA Averaged Weights(AG News, DBPedia)	60.64	79.31	80.50	78.12	50.58	71.67	31.25	48.83
Llama 3.1 8B	53.08	68.65	71.83	65.47	44.81	62.86	32.92	38.67
+ QLoRA AG News	**68.09**	**84.38**	**85.00**	83.76	**58.41**	**73.57**	36.92	**64.75**
+ QLoRA DBPedia	59.17	76.73	75.17	78.29	49.34	67.86	32.42	47.75
+ QLoRA Combined Topic dataset	66.31	82.60	83.50	81.71	56.35	70.48	**39.17**	59.42
+ QLoRA Output Summing(AG News, DBPedia)	64.68	83.64	82.83	**84.44**	53.45	71.19	30.50	58.67
+ QLoRA Output Averaging(AG News, DBPedia)	62.43	81.17	82.00	80.34	51.48	69.52	30.08	54.83
+ QLoRA Averaged Weights(AG News, DBPedia)	58.20	75.89	74.17	77.61	48.51	67.86	30.92	46.75
Mistral 7B	49.63	65.81	66.50	65.13	42.48	66.43	29.00	32.00
+ QLoRA AG News	**69.06**	**88.63**	**87.00**	**90.26**	**56.89**	73.10	**40.75**	**56.83**
+ QLoRA DBPedia	59.00	72.41	72.00	72.82	51.67	66.43	35.33	53.25
+ QLoRA Combined Topic dataset	65.04	82.62	82.33	82.91	54.39	70.00	37.08	56.08
+ QLoRA Output Summing(AG News, DBPedia)	66.01	83.98	83.00	84.96	56.08	**75.48**	39.75	53.00
+ QLoRA Output Averaging(AG News, DBPedia)	60.81	80.34	80.17	80.51	49.73	69.76	33.00	46.42
+ QLoRA Averaged Weights(AG News, DBPedia)	59.13	72.41	72.17	72.65	52.23	68.10	35.17	53.42

Table 7: Control Effectiveness calculated using an LLM as classifier for **Topic Control**, training on *single* and *combined* datasets, and composition of modules trained on single datasets Sum(data1, data2), Average(data1, data2), and Weights Average(data1, data2). Bold (shaded) = (second) highest score in column and section; underline = train and test set from same dataset.

C Weight Averaging vs Output Averaging

While both weight averaging and output averaging involve linear operations, they are *not* mathematically equivalent due to how the averaging interacts with the low-rank structure. An important distinction between weight averaging and output composition methods arises from our implementation: weight averaging averages the low-rank factors A and B separately ($A_{\text{avg}} = \frac{1}{N}\sum A_i$, $B_{\text{avg}} = \frac{1}{N}\sum B_i$), then computes the adapted output as $A_{\text{avg}}@B_{\text{avg}}$. This differs fundamentally from output averaging, which computes each module's full transformation $A_i@B_i$ before averaging.

When factors are averaged separately, the resulting computation includes cross-terms:

$$[\sum A_i]@[\sum B_j] = \sum_i (A_i B_i) + \sum_{i \neq j} (A_i B_j).$$

The cross-terms $A_i B_j$ $(i \neq j)$ represent combinations of the down-projection from module i with the up-projection from module j-components that were never trained to work together. These cross-terms are absent in output averaging, which only combines the trained transformations $A_i B_i$. For N modules, weight averaging produces N^2 terms while output averaging produces N terms. Additionally, averaging factors separately results in $1/N^2$ scaling, compared to $1/N$ for output averaging (or no normalization for output summing).

The presence of cross-terms means that weight averaging creates novel, untrained combinations of module components, which may explain why output composition methods consistently outperform weight averaging in our experiments (Tables 2-4).

D Models

In this study, we consider three different raw pre-trained LLMs: LLaMa 3 8B (AI@Meta, 2024), LLaMa 3.1 8B, and Mistral 7B (Jiang et al., 2023).

LLaMa 3 8B is a pre-trained autoregressive transformer model with 8 billion parameters. It has a decoder-only architecture and it is optimised with grouped-query attention for faster inference and SwiGLU activation function for improved training efficiency. Additionally, it incorporates Rotary Positional Embeddings (RoPE) to handle longer context lengths. The model was trained on 15T tokens that were collected from publicly available sources and highly curated to get a large high-quality training dataset.[3]

LLaMA 3.1 8B is an incremental update to LLaMA 3 8B, with the same 8 billion parameter transformer architecture but incorporating improvements in training stability, dataset quality, and tokenisation. While specific architectural details are not publicly disclosed, this version is expected to refine pretraining efficiency, dataset diversity, and

[3] https://ai.meta.com/blog/meta-llama-3/

Dataset	Pearson	Spearman
AG News	0.915 (0.000)	0.944 (0.000)
DBPedia	0.955 (0.000)	0.989 (0.000)
PPLM Prompt	0.917 (0.000)	0.928 (0.000)
STS	0.919 (0.000)	0.883 (0.000)
STS proc	0.941 (0.000)	0.944 (0.000)

Table 8: Pearson's and Spearman's correlations between the LLM as classifier and the other three classifiers used. p-value in parentheses.

representation quality, leading to better perplexity scores and more coherent text generation.

Mistral 7B is a pretrained transformer model with 7 billion parameters, designed to be highly efficient while maintaining strong language modeling performance. It incorporates grouped-query attention together with sliding window attention, which allows to efficiently handle longer context lengths without increasing computational complexity.

E Comparison of CE with LLM as classifier

Given that all our classifiers are trained on AG News, our evaluation setup might be biased favouring the texts generated by a system trained on AG News. To test whether this is the case or not, we performed an additional evaluation using a general-purpose instruction-tuned LLM as a classifier. We used Command R plus quantised in 4 bit,[4] using a simple prompt (Table 9) following Cohere's guidelines and template for classification.[5,6] The designed prompt is then formatted with the correct special tokens using the apply chat template function.

Our findings show that the overall trends, i.e. the settings identified as best or second-best, remain consistent with those observed in Table 3, indicating that the classifiers' training did not significantly distort our conclusions. To further validate this hypothesis, we computed Pearson's and Spearman's correlation coefficients between the Control Effectiveness scores obtained from the instruction-tuned LLM and those from the original classifiers.

The results (Table 8) show strong correlations, with Pearson's correlation ranging from 0.915 for AG News to 0.955 for DBPedia, and Spearman's correlation ranging from 0.883 for STS to 0.944 for AG News and STS proc. These high correlation values confirm that the instruction-tuned LLM behaves similarly to the classifiers originally used in our evaluation, reinforcing the reliability of our Control Effectiveness metric.

[4] https://huggingface.co/CohereForAI/c4ai-command-r-plus-4bit

[5] https://cohere.com/llmu/use-case-patterns#classifying

[6] https://huggingface.co/CohereForAI/c4ai-command-r-plus-4bit

F Module Composition Analysis

For a closer look at module composition not only preserving single-task performance, but even outperforming task-specialised single modules, Table 10 shows performance of models tested on individual test data sets (D_a) when trained on (just) D_a, compared to 2-module and 3-module compositions, for the sentiment control task.

The very clear trend is that the 3-module composition outperforms the others in 7 of 9 scenarios (average 93.33% vs. 91.44% for single-task modules), demonstrating (i) generalisation to the sentiment-control task generally (beyond individual datasets), and (ii) an advantage for single-dataset performance resulting from the generalisation. The two exceptions, L3-8B on SST-2 (-0.89%) and Mistral-7B on Yelp (-1.00%), represent marginal differences that do not change the core finding: output composition consistently matches or exceeds specialised module performance while enabling plug-and-play multi-task functionality.

G Standard Deviation Results

In all tables we report the results averaged across three runs with different seeds. We report all the results including standard deviation in Tables 11 and 12 for sentiment control, in Tables 13 and 14 for topic control, and in Tables 15 and 16 for multiple-attribute control.

H Hyperparameters

H.1 Trained Models

We train all our QLoRA modules with the following pre-trained raw models: `meta-llama/Meta-Llama-3-8B`, `meta-llama/Llama-3.1-8B`, and `mistralai/Mistral-7B-v0.3`. Each model is trained on 3 different seeds (8989, 79817, 794323). All the executions are done on a NVIDIA A100 with 80GB.

We perform a small grid search for each trained module investigating the learning rate (5e-6, 2e-4) and learning rate scheduler (cosine, constant).

System message:	You are a helpful assistant that classifies texts by topic.
User message:	Classify the following text into one of these topic categories: SPORTS, BUSINESS, WORLD, SCIENCE/TECHNOLOGY. Only reply with one of the possible topic categories. Do not include any other category or text. {text}

Table 9: Prompt used for the LLM as classifier.

Model	test	training / output summing			
	D_a	D_a	D_a, D_b	D_a, D_c	D_a, D_b, D_c
L3-8B	IMDB	89.67	89.33	89.56	**92.56**
	SST2	91.33	89.56	**93.22**	90.44
	Yelp	92.78	90.22	92.44	**94.44**
L3.1-8B	IMDB	91.44	90.00	90.89	**92.67**
	SST2	91.22	93.22	94.44	**95.11**
	Yelp	91.00	91.22	92.78	**96.33**
Mist7B	IMDB	88.56	85.44	86.33	**89.44**
	SST2	92.22	90.33	93.44	**95.22**
	Yelp	**94.78**	89.00	93.33	93.78
Average		91.44	89.81	91.83	**93.33**

Table 10: Sentiment control effectiveness (%) comparing single-dataset training with multi-module output summing. Bold=best; gray=second-best.

We use Paged AdamW (Loshchilov, 2017) as the optimiser.

We set all the other hyperparameters as *optimizer*=paged_adamw_32bit, *batch size*=4, *gradient accumulation steps*=4, *maximum gradient norm*=1.0, *warmup ratio*=0.1, *weight decay*=0.5, *group by length*=true, *fp16*=false, *bf16*=true, *maximum sequence length*=1024, *padding*=right, *save strategy*=epoch and we trained the modules for up to 3 epochs. During inference we set *batch size*=64.

Regarding model quantization, we set *bnb 4bit compute dtype*=bfloat16, *use 4bit*=True, *bnb 4bit quant type*=nf4, and *use nested quant*=False.

Regarding QLoRA hyperparameters, we follow the hyperparameters setting used in the vanilla QLoRA by Dettmers et al. (Appendix B.2). We set *rank*=64, *alpha*=16, *dropout*=0.1 and we attach QLoRA at every linear layer of the pre-trained model (["q_proj", "k_proj", "v_proj", "o_proj", "gate_proj", "up_proj", "down_proj", "lm_head"]).

To get the best trained module for each run, we evaluate Control Effectiveness of each saved checkpoint using the validation portion of the train dataset. We determine the best model for each grid search based on the run with highest Control Effectiveness. Tables 17 and 18 show the hyperamaters of the best-performing module for each executed grid search.

H.2 Evaluation Metrics

SLOR We used `gpt2-xl` and `bigscience/bloom-1b7` models from HuggingFace to compute sentence and unigram probabilities.

Control Effectiveness For sentiment control, we used `distilbert/distilbert-base-uncased-finetuned-sst-2-english`[7] and `michelecafagna26/t5-base-finetuned-sst2-sentiment`[8] from HuggingFace, and DeBERTa fine-tuned on Yelp (Gu et al., 2023). For topic control, we used `textattack/distilbert-base-uncased-ag-news`,[9] and `fabriceyhc/bert-base-uncased-ag_news`[10] from HuggingFace, and DeBERTa (Gu et al., 2023). For topic control, we further used a general-purpose intruction-tuned LLM as classifier, namely `CohereForAI/c4ai-command-r-plus-4bit`.[11]

I Scientific artifacts and licensing

Mistral 7B v0.3 and PPLM prompts are licensed under the Apache-2.0 license. LLaMa 3 8B[12] and LLaMa 3.1 8B[13] are licensed under a commercial license. Yelp Reviews dataset[14] is licensed under a commercial license. The DBPedia ontology classification dataset is licensed under the terms of the Creative Commons Attribution-ShareAlike License

[7] https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english

[8] https://huggingface.co/michelecafagna26/t5-base-finetuned-sst2-sentiment

[9] https://huggingface.co/textattack/distilbert-base-uncased-ag-news

[10] https://huggingface.co/fabriceyhc/bert-base-uncased-ag_news

[11] https://huggingface.co/CohereForAI/c4ai-command-r-plus-4bit

[12] https://huggingface.co/meta-llama/Meta-Llama-3-8B/blob/main/LICENSE

[13] https://huggingface.co/meta-llama/Llama-3.1-8B/blob/main/LICENSE

[14] https://s3-media3.fl.yelpcdn.com/assets/srv0/engineering_pages/bea5c1e92bf3/assets/vendor/yelp-dataset-agreement.pdf

CTG Technique	Distinct-n↑			SLOR↑
	dist-1	dist-2	dist-3	
Llama 3 8B	0.03± 0.01	0.09± 0.02	0.12± 0.03	9.44± 0.59
+ QLoRA Yelp	0.09± 0.04	0.24± 0.11	0.34± 0.15	9.87± 0.40
+ QLoRA IMDB	0.04± 0.02	0.13± 0.04	0.19± 0.06	**10.86**± 0.30
+ QLoRA SST-2	**0.34**± 0.08	**0.64**± 0.12	**0.71**± 0.12	8.24± 0.12
+ QLoRA Combined Sentiment dataset	0.11± 0.10	0.27± 0.20	0.35± 0.22	10.07± 0.95
+ QLoRA Output Summing(IMDB, SST-2)	0.15± 0.09	0.36± 0.19	0.46± 0.22	9.43± 0.39
+ QLoRA Output Summing(IMDB, Yelp)	0.07± 0.04	0.21± 0.09	0.31± 0.13	10.41± 0.38
+ QLoRA Output Summing(Yelp, SST-2)	0.21± 0.11	0.46± 0.22	0.56± 0.24	8.93± 0.34
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	0.13± 0.07	0.34± 0.18	0.45± 0.23	9.83± 0.50
+ QLoRA Output Averaging(IMDB, SST-2)	0.12± 0.07	0.27± 0.14	0.35± 0.16	9.49± 0.37
+ QLoRA Output Averaging(IMDB, Yelp)	0.05± 0.02	0.16± 0.06	0.23± 0.08	10.30± 0.25
+ QLoRA Output Averaging(Yelp, SST-2)	0.13± 0.08	0.31± 0.16	0.38± 0.19	9.16± 0.41
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	0.07± 0.04	0.19± 0.09	0.27± 0.12	9.77± 0.39
+ QLoRA Averaged Weights(IMDB, SST-2)	0.31± 0.11	0.61± 0.18	0.69± 0.17	8.37± 0.50
+ QLoRA Averaged Weights(IMDB, Yelp)	0.04± 0.02	0.12± 0.05	0.19± 0.06	10.83± 0.34
+ QLoRA Averaged Weights(Yelp, SST-2)	0.33± 0.08	**0.64**± 0.13	**0.71**± 0.12	8.28± 0.07
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	0.11± 0.06	0.28± 0.14	0.36± 0.17	9.54± 0.40
Llama 3.1 8B	0.04± 0.01	0.10± 0.03	0.13± 0.04	9.66± 0.63
+ QLoRA Yelp	0.09± 0.05	0.26± 0.13	0.36± 0.17	9.94± 0.50
+ QLoRA IMDB	0.04± 0.02	0.14± 0.05	0.21± 0.07	**10.81**± 0.33
+ QLoRA SST-2	**0.36**± 0.07	**0.69**± 0.11	**0.76**± 0.11	8.24± 0.11
+ QLoRA Combined Sentiment dataset	0.12± 0.11	0.27± 0.19	0.35± 0.21	10.00± 0.96
+ QLoRA Output Summing(IMDB, SST-2)	0.09± 0.04	0.23± 0.11	0.31± 0.14	9.91± 0.49
+ QLoRA Output Summing(IMDB, Yelp)	0.07± 0.04	0.21± 0.11	0.31± 0.16	10.33± 0.56
+ QLoRA Output Summing(Yelp, SST-2)	0.16± 0.09	0.38± 0.19	0.47± 0.23	9.16± 0.74
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	0.10± 0.06	0.27± 0.14	0.37± 0.19	9.91± 0.76
+ QLoRA Output Averaging(IMDB, SST-2)	0.08± 0.04	0.21± 0.09	0.28± 0.11	9.81± 0.58
+ QLoRA Output Averaging(IMDB, Yelp)	0.07± 0.04	0.22± 0.10	0.31± 0.14	10.24± 0.34
+ QLoRA Output Averaging(Yelp, SST-2)	0.16± 0.10	0.36± 0.20	0.44± 0.23	9.05± 0.59
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	0.07± 0.04	0.18± 0.08	0.26± 0.11	10.01± 0.60
+ QLoRA Averaged Weights(IMDB, SST-2)	0.33± 0.12	0.64± 0.21	0.71± 0.20	8.38± 0.61
+ QLoRA Averaged Weights(IMDB, Yelp)	0.04± 0.03	0.14± 0.06	0.21± 0.08	10.78± 0.42
+ QLoRA Averaged Weights(Yelp, SST-2)	0.35± 0.08	0.67± 0.12	0.74± 0.11	8.27± 0.12
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	0.08± 0.04	0.20± 0.09	0.27± 0.12	9.87± 0.59
Mistral 7B	0.04± 0.02	0.09± 0.03	0.12± 0.03	7.45± 1.72
+ QLoRA Yelp	0.03± 0.01	0.07± 0.03	0.10± 0.04	10.44± 0.68
+ QLoRA IMDB	0.02± 0.01	0.04± 0.02	0.06± 0.03	**11.25**± 0.95
+ QLoRA SST-2	**0.27**± 0.06	**0.51**± 0.10	**0.58**± 0.10	7.65± 0.19
+ QLoRA Combined Sentiment dataset	0.03± 0.03	0.06± 0.05	0.09± 0.05	10.83± 1.13
+ QLoRA Output Summing(IMDB, SST-2)	0.22± 0.09	0.41± 0.15	0.47± 0.15	7.80± 0.47
+ QLoRA Output Summing(IMDB, Yelp)	0.03± 0.01	0.07± 0.03	0.10± 0.04	10.51± 0.68
+ QLoRA Output Summing(Yelp, SST-2)	0.18± 0.10	0.35± 0.17	0.40± 0.20	8.22± 1.05
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	0.08± 0.08	0.16± 0.13	0.19± 0.13	9.13± 1.12
+ QLoRA Output Averaging(IMDB, SST-2)	0.11± 0.07	0.21± 0.11	0.25± 0.12	8.49± 0.56
+ QLoRA Output Averaging(IMDB, Yelp)	0.02± 0.01	0.06± 0.03	0.09± 0.04	10.80± 0.72
+ QLoRA Output Averaging(Yelp, SST-2)	0.05± 0.03	0.10± 0.06	0.13± 0.06	8.84± 0.61
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	0.02± 0.01	0.05± 0.02	0.07± 0.02	10.54± 0.75
+ QLoRA Averaged Weights(IMDB, SST-2)	0.25± 0.06	0.49± 0.11	0.55± 0.11	7.69± 0.25
+ QLoRA Averaged Weights(IMDB, Yelp)	0.02± 0.01	0.04± 0.02	0.07± 0.03	11.19± 0.93
+ QLoRA Averaged Weights(Yelp, SST-2)	**0.27**± 0.05	0.50± 0.08	0.57± 0.09	7.66± 0.19
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	0.10± 0.07	0.20± 0.12	0.24± 0.12	8.64± 1.07

Table 11: **Sentiment Control** Diversity, Fluency for the model + QLoRA module combinations explained in Section 4. Here, e.g. Output Summing(data1, data2) refers to the output summation module composition technique. All values are averages over 3 runs and standard deviation is reported. Bold (shaded) = (second) highest score in column/section.

and the GNU Free Documentation License. IMDB, SST-2, and AG News datasets do not specify a license. STS benchmark test provides a license for each included dataset.[15] regarding the classifiers used in the evaluation, DistilBERT and T5 fine-tuned on SST-2 are licensed under the Apach -2.0 license. BERT fine-tuned on AG News is licensed under the Apache-2.0 license, while DeBERTa models fine-tuned by Gu et al. (2023) are licensed under the MIT license. Command R plus is licensed under the Creative Commons Attribution Non Commercial 4.0. The usage of all listed artifacts is consistent with their licenses.

[15] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark

CTG Technique	Control Effectiveness↑								
	Avg All	Avg	Yelp	IMDB	SST-2	Out-Of-Domain			
						Avg	PPLM S	STS S	STS proc S
Llama 3 8B	60.43	63.70	68.00± 2.03	57.56± 3.37	65.56± 1.68	57.16	64.60± 3.97	52.83± 0.88	54.06± 2.06
+ QLoRA Yelp	85.47	91.44	92.78± 1.17	90.67± 3.48	90.89± 0.51	79.49	88.41± 1.92	**67.78**± 0.69	82.28± 1.29
+ QLoRA IMDB	82.84	92.07	92.11± 1.07	89.67± 1.33	94.44± 0.19	73.61	80.95± 1.65	62.00± 1.96	77.89± 2.44
+ QLoRA SST-2	78.35	85.96	84.11± 3.47	82.44± 2.71	91.33± 2.40	70.73	82.86± 5.85	57.39± 2.01	71.94± 1.78
+ QLoRA Combined Sentiment dataset	83.47	91.07	92.22± 2.41	92.00± 3.18	89.00± 3.71	75.88	87.46± 2.15	63.50± 2.19	76.67± 2.84
+ QLoRA Output Summing(IMDB, SST-2)	82.85	90.85	89.78± 3.75	89.56± 0.51	93.22± 1.64	74.85	88.89± 4.16	61.72± 4.19	73.94± 2.26
+ QLoRA Output Summing(IMDB, Yelp)	84.35	91.22	90.22± 0.84	89.33± 2.03	94.11± 1.17	77.49	87.46± 1.92	64.44± 0.98	80.56± 1.68
+ QLoRA Output Summing(Yelp, SST-2)	82.85	90.74	92.44± 1.07	90.22± 2.87	89.56± 1.68	74.96	**90.00**± 3.12	58.33± 1.69	76.56± 3.13
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	**86.01**	**92.48**	**94.44**± 1.71	**92.56**± 0.84	90.44± 1.71	**79.54**	88.89± 0.99	67.00± 3.32	**82.72**± 2.64
+ QLoRA Output Averaging(IMDB, SST-2)	79.77	88.30	86.33± 0.67	87.44± 1.02	91.11± 1.92	71.24	83.65± 3.17	58.44± 2.34	71.61± 0.92
+ QLoRA Output Averaging(IMDB, Yelp)	83.97	92.26	91.11± 2.27	89.67± 4.16	**96.00**± 0.88	75.67	86.19± 4.59	63.17± 3.62	77.67± 4.67
+ QLoRA Output Averaging(Yelp, SST-2)	82.21	89.74	89.00± 1.73	88.56± 1.90	91.67± 2.65	74.69	84.29± 2.90	62.61± 2.27	77.17± 1.73
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	81.42	89.81	88.78± 0.84	89.00± 3.06	91.67± 1.15	73.03	84.60± 2.44	60.50± 0.83	74.00± 0.29
+ QLoRA Averaged Weights(IMDB, SST-2)	77.54	85.93	85.33± 2.08	81.89± 1.58	90.56± 0.69	69.16	77.94± 1.80	57.50± 2.40	72.06± 2.55
+ QLoRA Averaged Weights(IMDB, Yelp)	82.20	90.74	91.89± 0.84	87.22± 2.41	93.11± 0.69	73.65	84.13± 5.05	60.67± 1.61	76.17± 1.17
+ QLoRA Averaged Weights(Yelp, SST-2)	78.68	85.93	85.33± 2.08	81.89± 1.58	90.56± 0.69	71.44	84.76± 1.43	57.50± 2.40	72.06± 2.55
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	80.60	90.04	88.89± 2.34	88.78± 1.90	92.44± 0.96	71.17	83.17± 2.15	58.72± 2.08	71.61± 0.51
Llama 3.1 8B	58.96	61.52	66.22± 2.34	54.22± 1.17	64.11± 1.35	56.41	64.29± 3.33	51.56± 2.41	53.39± 3.51
+ QLoRA Yelp	87.23	91.52	91.00± 2.65	89.00± 2.65	94.56± 1.84	**82.93**	91.75± 3.17	**71.28**± 1.60	**85.78**± 1.86
+ QLoRA IMDB	83.65	92.26	89.67± 3.28	91.44± 1.17	**95.67**± 1.20	75.04	85.24± 1.26	63.33± 0.29	76.56± 3.19
+ QLoRA SST-2	78.99	85.48	82.33± 1.45	82.89± 2.78	91.22± 1.92	72.49	84.76± 4.54	58.17± 1.76	74.56± 0.25
+ QLoRA Combined Sentiment dataset	85.26	92.67	94.89± 2.04	91.44± 0.51	91.67± 1.20	77.86	89.52± 4.59	62.28± 1.42	81.78± 3.14
+ QLoRA Output Summing(IMDB, SST-2)	85.80	92.93	93.44± 1.35	90.89± 4.55	94.44± 0.69	78.68	**92.54**± 1.67	62.78± 0.42	80.72± 1.40
+ QLoRA Output Summing(IMDB, Yelp)	85.89	91.52	91.22± 0.19	90.00± 5.51	93.33± 2.85	80.26	89.05± 1.72	68.06± 0.86	83.67± 1.48
+ QLoRA Output Summing(Yelp, SST-2)	85.78	92.48	92.78± 1.17	91.44± 2.83	93.22± 0.38	79.07	89.05± 2.08	68.94± 0.59	79.22± 1.00
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	**87.66**	**94.70**	**96.33**± 1.33	**92.67**± 1.20	95.11± 1.71	80.61	90.16± 0.27	66.89± 3.47	84.78± 2.06
+ QLoRA Output Averaging(IMDB, SST-2)	81.35	89.74	89.22± 3.47	87.22± 1.26	92.78± 1.17	72.95	83.02± 0.99	61.50± 1.83	74.33± 2.62
+ QLoRA Output Averaging(IMDB, Yelp)	83.73	92.30	93.00± 1.76	90.22± 3.36	93.67± 1.76	75.16	86.03± 2.79	63.22± 2.41	76.22± 3.15
+ QLoRA Output Averaging(Yelp, SST-2)	81.03	89.00	89.67± 4.91	86.89± 2.27	90.44± 1.35	73.06	85.08± 2.25	60.94± 2.56	73.17± 1.42
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	80.42	89.15	88.56± 2.12	87.11± 1.50	91.78± 0.19	71.69	85.56± 2.62	58.39± 1.73	71.11± 3.64
+ QLoRA Averaged Weights(IMDB, SST-2)	77.64	84.07	81.00± 3.67	81.89± 3.02	89.33± 5.13	71.21	83.97± 1.92	56.67± 2.17	73.00± 1.04
+ QLoRA Averaged Weights(IMDB, Yelp)	82.38	91.19	90.67± 2.19	88.89± 3.66	94.00± 1.86	73.57	83.49± 3.17	61.89± 4.03	75.33± 1.74
+ QLoRA Averaged Weights(Yelp, SST-2)	77.54	84.07	81.00± 3.67	81.89± 3.02	89.33± 5.13	71.00	83.33± 1.65	56.67± 2.17	73.00± 1.04
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	81.93	89.67	91.67± 0.58	86.67± 1.67	90.67± 1.45	74.19	85.56± 2.25	60.72± 0.63	76.28± 1.55
Mistral 7B	58.12	60.00	59.33± 0.00	55.00± 0.00	65.67± 0.00	56.23	62.86± 0.00	52.50± 0.00	53.33± 0.00
+ QLoRA Yelp	**84.36**	92.30	**94.78**± 3.56	**90.56**± 2.71	91.56± 2.04	**76.42**	84.60± 0.73	**61.33**± 1.48	**83.33**± 1.20
+ QLoRA IMDB	80.51	90.63	89.67± 0.33	88.56± 4.40	93.67± 2.96	70.38	82.70± 2.20	58.78± 1.25	69.67± 3.92
+ QLoRA SST-2	78.77	85.67	83.11± 1.02	81.67± 1.67	92.22± 1.07	71.87	83.65± 6.12	55.94± 0.92	76.00± 1.64
+ QLoRA Combined Sentiment dataset	78.81	86.30	87.11± 0.69	85.67± 5.70	86.11± 3.75	71.32	83.17± 3.24	58.22± 0.95	72.56± 2.43
+ QLoRA Output Summing(IMDB, SST-2)	80.12	87.11	85.56± 2.55	85.44± 1.50	90.33± 3.33	73.12	86.98± 3.57	57.11± 1.78	75.28± 5.09
+ QLoRA Output Summing(IMDB, Yelp)	83.02	90.44	93.33± 2.73	86.33± 3.71	91.67± 4.10	75.59	86.98± 2.44	58.28± 1.00	81.50± 2.33
+ QLoRA Output Summing(Yelp, SST-2)	81.20	88.26	89.00± 0.58	82.33± 1.15	93.44± 1.71	74.13	**89.68**± 1.67	59.22± 1.51	73.50± 3.22
+ QLoRA Output Summing(IMDB, Yelp, SST-2)	83.45	**92.81**	93.78± 1.95	89.44± 1.26	**95.22**± 2.22	74.09	86.67± 1.26	58.89± 0.84	76.72± 2.18
+ QLoRA Output Averaging(IMDB, SST-2)	79.41	88.89	88.56± 1.07	85.33± 1.53	92.78± 1.02	69.93	82.06± 0.27	58.39± 0.42	69.33± 2.62
+ QLoRA Output Averaging(IMDB, Yelp)	82.49	92.63	93.56± 1.07	90.11± 1.02	94.22± 0.77	72.35	84.60± 2.75	56.78± 0.35	75.67± 0.83
+ QLoRA Output Averaging(Yelp, SST-2)	81.35	90.74	92.44± 1.54	86.33± 3.84	93.44± 0.84	71.96	85.87± 0.99	60.67± 0.73	69.33± 2.25
+ QLoRA Output Averaging(IMDB, Yelp, SST-2)	81.68	92.37	91.67± 0.67	**90.56**± 2.71	94.89± 0.51	70.99	84.76± 4.15	57.67± 2.18	70.56± 0.84
+ QLoRA Averaged Weights(IMDB, SST-2)	78.92	85.85	82.89± 1.17	82.22± 2.01	92.44± 1.50	71.98	84.60± 2.91	56.28± 0.51	75.06± 3.23
+ QLoRA Averaged Weights(IMDB, Yelp)	81.03	91.33	90.22± 1.71	89.89± 4.86	93.89± 2.78	70.72	83.65± 2.25	59.56± 1.84	68.94± 2.51
+ QLoRA Averaged Weights(Yelp, SST-2)	79.23	85.85	82.89± 1.17	82.22± 2.01	92.44± 1.50	72.61	86.51± 4.50	56.28± 0.51	75.06± 3.23
+ QLoRA Averaged Weights(IMDB, Yelp, SST-2)	79.97	88.26	88.11± 2.27	84.11± 0.69	92.56± 0.77	71.67	86.35± 3.10	58.00± 0.44	70.67± 2.84

Table 12: **Sentiment Control** Control Effectiveness for the model + QLoRA module combinations explained in Section 4. Here, e.g. Output Summing(data1, data2) refers to the output summation module composition technique. All values are averages over 3 runs and standard deviation is reported for single datasets results. Bold (shaded) = (second) highest score in column/section; underline = train/test on same dataset.

CTG Technique	Distinct-n↑			SLOR↑
	dist-1	dist-2	dist-3	
Llama 3 8B	0.07± 0.02	0.17± 0.05	0.22± 0.05	8.45± 0.74
+ QLoRA AG News	0.25± 0.09	0.53± 0.15	0.62± 0.16	9.39± 0.13
+ QLoRA DBPedia	0.32± 0.09	0.60± 0.13	0.68± 0.13	8.96± 0.37
+ QLoRA Combined Topic dataset	0.30± 0.10	0.60± 0.15	0.70± 0.15	8.92± 0.30
+ QLoRA Output Summing(AG News, DBPedia)	**0.35**± 0.08	**0.70**± 0.11	**0.80**± 0.09	9.33± 0.36
+ QLoRA Output Averaging(AG News, DBPedia)	0.27± 0.10	0.55± 0.17	0.64± 0.18	**9.57**± 0.17
+ QLoRA Averaged Weights(AG News, DBPedia)	0.33± 0.08	0.62± 0.11	0.71± 0.12	8.98± 0.41
Llama 3.1 8B	0.05± 0.02	0.12± 0.04	0.17± 0.03	**9.45**± 0.76
+ QLoRA AG News	0.26± 0.09	0.55± 0.15	0.65± 0.15	9.40± 0.11
+ QLoRA DBPedia	0.31± 0.08	0.59± 0.11	0.68± 0.11	8.74± 0.45
+ QLoRA Combined Topic dataset	0.31± 0.10	0.62± 0.16	0.73± 0.16	9.29± 0.20
+ QLoRA Output Summing(AG News, DBPedia)	**0.35**± 0.07	**0.68**± 0.10	**0.78**± 0.09	9.12± 0.39
+ QLoRA Output Averaging(AG News, DBPedia)	0.30± 0.09	0.59± 0.14	0.68± 0.14	8.98± 0.23
+ QLoRA Averaged Weights(AG News, DBPedia)	0.32± 0.08	0.61± 0.11	0.71± 0.10	8.74± 0.37
Mistral 7B	0.07± 0.05	0.13± 0.07	0.16± 0.06	7.16± 2.31
+ QLoRA AG News	0.18± 0.10	0.37± 0.17	0.44± 0.18	9.48± 0.28
+ QLoRA DBPedia	**0.26**± 0.06	0.48± 0.08	0.56± 0.08	8.81± 0.54
+ QLoRA Combined Topic dataset	0.21± 0.09	0.43± 0.14	0.52± 0.15	9.09± 0.31
+ QLoRA Output Summing(AG News, DBPedia)	**0.26**± 0.09	**0.50**± 0.15	**0.59**± 0.15	9.26± 0.17
+ QLoRA Output Averaging(AG News, DBPedia)	0.18± 0.10	0.36± 0.16	0.42± 0.18	**9.60**± 0.13
+ QLoRA Averaged Weights(AG News, DBPedia)	**0.26**± 0.06	0.48± 0.08	0.56± 0.08	8.87± 0.62

Table 13: Diversity, Fluency for **Topic Control**, training on *single* and *combined* datasets, and composition of modules trained on single datasets, e.g. Output Summing(data1, data2). All values are averages over 3 runs and standard deviation is reported. Bold (shaded) = (second) highest score in column and section.

CTG Technique	Control Effectiveness↑							
	Avg All	Avg	AG News	DBPedia	Out-Of-Domain			
					Avg	PPLM T	STS T	STS proc T
Llama 3 8B	45.92	58.52	64.61± 2.43	52.42± 2.32	37.53	48.97± 3.49	27.97± 1.35	35.64± 1.25
+ QLoRA AG News	**68.86**	**85.13**	**90.72**± 0.42	**79.54**± 0.99	**58.02**	**71.11**± 4.33	**42.03**± 0.39	60.92± 2.21
+ QLoRA DBPedia	52.97	69.94	71.67± 1.61	68.21± 2.22	41.66	55.40± 1.45	28.58± 1.63	41.00± 1.80
+ QLoRA Combined Topic dataset	63.53	74.42	81.61± 1.11	67.24± 1.45	56.27	68.73± 4.73	37.50± 2.11	**62.58**± 0.17
+ QLoRA Output Summing(AG News, DBPedia)	64.58	82.57	88.78± 0.35	76.35± 0.43	52.59	**71.11**± 1.31	32.81± 0.43	53.86± 1.24
+ QLoRA Output Averaging(AG News, DBPedia)	61.75	78.33	83.78± 2.17	72.88± 0.94	50.70	66.98± 1.59	33.81± 1.95	51.31± 1.85
+ QLoRA Averaged Weights(AG News, DBPedia)	55.55	70.61	72.11± 1.92	69.12± 3.08	45.50	66.59± 0.90	28.06± 1.85	41.86± 1.64
Llama 3.1 8B	45.93	58.61	66.11± 1.21	51.11± 2.52	37.47	46.35± 2.21	29.89± 0.49	36.17± 0.36
+ QLoRA AG News	**68.53**	**85.52**	**91.33**± 1.50	**79.72**± 0.36	**57.21**	**73.10**± 2.58	36.28± 0.34	**62.25**± 0.66
+ QLoRA DBPedia	52.93	69.86	70.94± 0.92	68.77± 3.27	41.64	54.92± 2.55	28.11± 1.10	41.89± 1.97
+ QLoRA Combined Topic dataset	65.27	80.65	89.06± 0.69	72.25± 2.23	55.01	66.51± 1.37	**39.97**± 6.45	58.56± 5.01
+ QLoRA Output Summing(AG News, DBPedia)	64.15	82.91	88.39± 1.36	77.44± 0.89	51.64	69.05± 2.42	29.39± 2.19	56.47± 0.97
+ QLoRA Output Averaging(AG News, DBPedia)	59.21	76.42	83.44± 2.07	69.40± 1.04	47.74	61.67± 1.09	29.81± 2.16	51.75± 1.53
+ QLoRA Averaged Weights(AG News, DBPedia)	53.64	69.52	70.94± 1.00	68.09± 2.57	43.06	60.71± 0.95	27.61± 0.64	40.86± 1.47
Mistral 7B	44.87	55.96	62.00± 0.00	49.91± 0.00	37.48	49.76± 0.00	29.92± 0.00	32.75± 0.00
+ QLoRA AG News	**69.69**	**88.52**	**93.17**± 0.88	**83.87**± 1.38	**57.14**	**73.97**± 2.50	**39.69**± 2.71	**57.75**± 0.80
+ QLoRA DBPedia	53.08	67.05	67.50± 1.04	66.61± 0.94	43.76	54.13± 0.50	32.89± 0.39	44.28± 1.84
+ QLoRA Combined Topic dataset	65.02	81.74	89.00± 2.35	74.47± 1.11	53.87	67.54± 3.86	39.22± 2.88	54.86± 4.65
+ QLoRA Output Summing(AG News, DBPedia)	66.02	85.27	88.89± 0.54	81.65± 1.31	53.19	69.76± 2.18	36.36± 0.69	53.44± 2.35
+ QLoRA Output Averaging(AG News, DBPedia)	58.39	76.01	82.00± 3.53	70.03± 0.94	46.64	60.95± 1.09	33.89± 0.94	45.08± 0.55
+ QLoRA Averaged Weights(AG News, DBPedia)	54.65	66.53	66.06± 1.17	67.01± 0.51	46.72	61.75± 1.76	33.47± 0.98	44.94± 2.96

Table 14: Diversity, Fluency, Control Effectiveness for **Topic Control**, training on *single* and *combined* datasets, and composition of modules trained on single datasets, e.g. Output Summing(data1, data2). All values are averages over 3 runs and standard deviation is reported for single dataset results. Bold (shaded) = (second) highest score in column and section; underline = train and test set from same dataset.

CTG Technique	Distinct-n↑			SLOR↑
	dist1	dist2	dist3	
Llama 3 8B	0.03 ± 0.00	0.10 ± 0.01	0.14 ± 0.01	8.36 ± 0.11
+ QLoRA Combined S	0.03 ± 0.00	0.09 ± 0.01	0.14 ± 0.01	**10.81 ± 0.12**
+ QLoRA Combined T	**0.24 ± 0.02**	**0.50 ± 0.04**	**0.59 ± 0.05**	8.52 ± 0.20
+ QLoRA Sum(Ind mod)	0.18 ± 0.01	0.42 ± 0.03	0.51 ± 0.04	9.11 ± 0.07
+ QLoRA Sum(S, T)	0.09 ± 0.01	0.23 ± 0.02	0.31 ± 0.03	9.81 ± 0.24
+ QLoRA Average(Ind mod)	0.08 ± 0.01	0.20 ± 0.01	0.26 ± 0.01	9.55 ± 0.05
+ QLoRA Average(S, T)	0.10 ± 0.03	0.24 ± 0.07	0.31 ± 0.08	9.26 ± 0.21
+ QLoRA Weights Average(Ind mod)	0.09 ± 0.01	0.23 ± 0.03	0.30 ± 0.03	9.45 ± 0.10
+ QLoRA Weights Average(S, T)	0.21 ± 0.04	0.46 ± 0.07	0.55 ± 0.07	8.56 ± 0.24
Llama 3.1 8B	0.05 ± 0.00	0.12 ± 0.01	0.18 ± 0.01	9.84 ± 0.09
+ QLoRA Combined S	0.03 ± 0.01	0.11 ± 0.02	0.17 ± 0.03	**10.73 ± 0.17**
+ QLoRA Combined T	**0.28 ± 0.05**	**0.58 ± 0.07**	**0.68 ± 0.08**	8.58 ± 0.50
+ QLoRA Sum(Ind mod)	0.13 ± 0.02	0.29 ± 0.05	0.35 ± 0.06	8.63 ± 0.16
+ QLoRA Sum(S, T)	0.10 ± 0.07	0.27 ± 0.12	0.36 ± 0.13	9.80 ± 0.49
+ QLoRA Average(Ind mod)	0.08 ± 0.01	0.22 ± 0.02	0.29 ± 0.02	9.36 ± 0.08
+ QLoRA Average(S, T)	0.11 ± 0.07	0.29 ± 0.13	0.37 ± 0.15	9.67 ± 0.57
+ QLoRA Weights Average(Ind mod)	0.10 ± 0.01	0.25 ± 0.03	0.33 ± 0.03	9.14 ± 0.04
+ QLoRA Weights Average(S, T)	0.24 ± 0.07	0.52 ± 0.13	0.62 ± 0.15	8.74 ± 0.63
Mistral 7B	0.03 ± 0.00	0.07 ± 0.00	0.10 ± 0.00	9.77 ± 0.00
+ QLoRA Combined S	0.01 ± 0.00	0.03 ± 0.00	0.05 ± 0.00	**11.60 ± 0.11**
+ QLoRA Combined T	**0.12 ± 0.03**	**0.27 ± 0.08**	**0.36 ± 0.11**	8.92 ± 0.31
+ QLoRA Sum(Ind mod)	0.09 ± 0.06	0.20 ± 0.10	0.24 ± 0.11	8.94 ± 0.54
+ QLoRA Sum(S, T)	0.09 ± 0.07	0.22 ± 0.16	0.30 ± 0.21	9.50 ± 1.23
+ QLoRA Average(Ind mod)	0.02 ± 0.00	0.05 ± 0.00	0.08 ± 0.00	10.75 ± 0.20
+ QLoRA Average(S, T)	0.08 ± 0.07	0.18 ± 0.14	0.23 ± 0.16	9.63 ± 1.01
+ QLoRA Weights Average(Ind mod)	0.03 ± 0.00	0.07 ± 0.01	0.10 ± 0.01	10.18 ± 0.35
+ QLoRA Weights Average(S, T)	0.10 ± 0.03	0.24 ± 0.08	0.32 ± 0.11	9.04 ± 0.48

Table 15: Diversity, Fluency **Multi-attribute Control** alongside single-attribute control results for comparison. All values are averages over 3 runs and standard deviation is reported. S=the Combined Sentiment dataset, T=the Combined Topic dataset, Ind mod=composition is on all 5 individually trained modules. Bold (shaded) = (second) highest score in column and section; underline = train and test set from same dataset.

CTG Technique	Control Effectiveness ↑									
	Out-Of-Domain									
	Multiple				Sentiment			Topic		
	Avg	PPLM M	STS M	STS p M	PPLM S	STS S	STS p S	PPLM T	STS T	STS p T
Llama 3 8B	19.22	29.29 ± 0.02	14.54 ± 0.00	20.38 ± 0.01	0.65 ± 0.04	0.53 ± 0.01	0.54 ± 0.02	0.49 ± 0.03	0.28 ± 0.01	0.36 ± 0.01
+ QLoRA Combined S	**25.66**	37.14 ± 0.03	16.50 ± 0.01	**30.79 ± 0.02**	0.87 ± 0.02	**0.64 ± 0.02**	**0.77 ± 0.03**	0.52 ± 0.02	0.28 ± 0.02	0.38 ± 0.01
+ QLoRA Combined T	22.64	36.55 ± 0.03	14.96 ± 0.01	25.46 ± 0.00	0.56 ± 0.02	0.51 ± 0.00	0.55 ± 0.02	**0.69 ± 0.05**	**0.38 ± 0.02**	**0.63 ± 0.00**
+ QLoRA Sum(Ind mod)	24.08	33.69 ± 0.02	16.79 ± 0.01	28.00 ± 0.01	0.80 ± 0.05	0.59 ± 0.05	0.69 ± 0.01	0.44 ± 0.06	0.25 ± 0.01	0.38 ± 0.05
+ QLoRA Sum(S, T)	23.85	35.60 ± 0.03	**17.75 ± 0.01**	25.83 ± 0.02	**0.88 ± 0.04**	0.60 ± 0.01	0.73 ± 0.03	0.62 ± 0.07	0.29 ± 0.02	0.50 ± 0.02
+ QLoRA Average(Ind mod)	23.30	**37.50 ± 0.03**	15.46 ± 0.01	26.17 ± 0.00	0.78 ± 0.04	0.57 ± 0.01	0.68 ± 0.05	0.58 ± 0.02	0.27 ± 0.01	0.43 ± 0.01
+ QLoRA Average(S, T)	22.32	36.55 ± 0.03	14.50 ± 0.01	25.17 ± 0.02	0.76 ± 0.07	0.55 ± 0.01	0.64 ± 0.03	0.58 ± 0.05	0.29 ± 0.02	0.42 ± 0.01
+ QLoRA Weights Average(Ind mod)	23.58	35.71 ± 0.04	15.79 ± 0.01	27.12 ± 0.01	0.74 ± 0.03	0.54 ± 0.03	0.63 ± 0.03	0.65 ± 0.02	0.32 ± 0.02	0.46 ± 0.02
+ QLoRA Weights Average(S, T)	22.85	36.55 ± 0.04	15.62 ± 0.01	25.29 ± 0.00	0.75 ± 0.01	0.51 ± 0.00	0.55 ± 0.02	0.68 ± 0.03	0.37 ± 0.02	0.60 ± 0.02
Llama 3.1 8B	19.95	28.57 ± 0.02	15.00 ± 0.01	21.88 ± 0.01	0.64 ± 0.03	0.52 ± 0.02	0.53 ± 0.04	0.46 ± 0.02	0.30 ± 0.00	0.36 ± 0.00
+ QLoRA Combined S	26.44	37.50 ± 0.04	17.54 ± 0.01	**31.46 ± 0.01**	**0.90 ± 0.05**	0.62 ± 0.01	**0.82 ± 0.03**	0.49 ± 0.07	0.31 ± 0.01	0.37 ± 0.01
+ QLoRA Combined T	23.95	35.12 ± 0.05	17.54 ± 0.07	26.46 ± 0.07	0.52 ± 0.05	0.51 ± 0.02	0.52 ± 0.02	0.67 ± 0.01	**0.40 ± 0.06**	0.59 ± 0.05
+ QLoRA Sum(Ind mod)	20.78	30.24 ± 0.03	16.29 ± 0.01	21.96 ± 0.02	0.85 ± 0.03	0.61 ± 0.02	0.73 ± 0.01	0.35 ± 0.04	0.24 ± 0.01	0.29 ± 0.02
+ QLoRA Sum(S, T)	**26.76**	34.40 ± 0.03	**19.58 ± 0.03**	31.25 ± 0.03	0.76 ± 0.20	**0.63 ± 0.05**	0.72 ± 0.13	0.58 ± 0.14	0.35 ± 0.11	0.50 ± 0.15
+ QLoRA Average(Ind mod)	23.17	37.38 ± 0.03	15.21 ± 0.00	26.17 ± 0.00	0.70 ± 0.01	0.55 ± 0.02	0.65 ± 0.02	0.46 ± 0.01	0.27 ± 0.01	0.39 ± 0.00
+ QLoRA Average(S, T)	25.89	**40.83 ± 0.07**	16.50 ± 0.02	30.04 ± 0.01	0.74 ± 0.12	0.60 ± 0.05	0.64 ± 0.09	0.59 ± 0.04	0.30 ± 0.04	0.43 ± 0.09
+ QLoRA Weights Average(Ind mod)	23.26	40.36 ± 0.04	15.46 ± 0.01	25.08 ± 0.01	0.72 ± 0.04	0.54 ± 0.03	0.64 ± 0.01	0.50 ± 0.02	0.28 ± 0.01	0.40 ± 0.02
+ QLoRA Weights Average(S, T)	24.43	37.98 ± 0.06	17.29 ± 0.07	26.83 ± 0.05	0.73 ± 0.14	0.51 ± 0.02	0.52 ± 0.02	**0.70 ± 0.03**	**0.40 ± 0.07**	**0.61 ± 0.05**
Mistral 7B	20.43	26.07 ± 0.00	16.12 ± 0.00	22.75 ± 0.00	0.63 ± 0.00	0.53 ± 0.00	0.53 ± 0.00	0.50 ± 0.00	0.30 ± 0.00	0.33 ± 0.00
+ QLoRA Combined S	21.47	30.60 ± 0.01	14.75 ± 0.01	25.00 ± 0.01	0.83 ± 0.03	**0.58 ± 0.01**	0.73 ± 0.02	0.47 ± 0.04	0.27 ± 0.01	0.38 ± 0.02
+ QLoRA Combined T	26.06	35.12 ± 0.04	20.33 ± 0.03	28.62 ± 0.03	0.50 ± 0.03	0.52 ± 0.01	0.52 ± 0.01	0.68 ± 0.04	0.39 ± 0.03	0.55 ± 0.05
+ QLoRA Sum(Ind mod)	22.77	36.19 ± 0.03	16.25 ± 0.01	24.58 ± 0.02	**0.88 ± 0.02**	0.57 ± 0.03	**0.76 ± 0.02**	0.57 ± 0.04	0.31 ± 0.01	0.42 ± 0.02
+ QLoRA Sum(S, T)	25.53	34.05 ± 0.03	19.58 ± 0.04	28.50 ± 0.03	0.60 ± 0.15	0.51 ± 0.05	0.58 ± 0.12	0.63 ± 0.01	**0.40 ± 0.01**	**0.61 ± 0.01**
+ QLoRA Average(Ind mod)	23.01	35.95 ± 0.01	14.08 ± 0.01	27.42 ± 0.02	0.76 ± 0.02	0.55 ± 0.02	0.67 ± 0.03	0.48 ± 0.03	0.28 ± 0.01	0.38 ± 0.01
+ QLoRA Average(S, T)	22.73	34.88 ± 0.07	15.54 ± 0.02	25.67 ± 0.02	0.63 ± 0.11	0.53 ± 0.05	0.57 ± 0.06	0.61 ± 0.03	0.30 ± 0.00	0.44 ± 0.04
+ QLoRA Weights Average(Ind mod)	21.70	**36.31 ± 0.01**	14.42 ± 0.00	23.88 ± 0.02	0.77 ± 0.01	0.54 ± 0.00	0.62 ± 0.00	0.53 ± 0.01	0.28 ± 0.00	0.39 ± 0.01
+ QLoRA Weights Average(S, T)	**26.24**	27.14 ± 0.10	**21.25 ± 0.04**	**30.92 ± 0.04**	0.59 ± 0.15	0.52 ± 0.01	0.52 ± 0.01	**0.70 ± 0.01**	**0.40 ± 0.03**	0.56 ± 0.05

Table 16: Control Effectiveness for **Multi-attribute Control** alongside single-attribute control results for comparison. All values are averages over 3 runs and standard deviation is reported for single dataset results. S=the Combined Sentiment dataset, T=the Combined Topic dataset, Ind mod=composition is on all 5 individually trained modules. Bold (shaded) = (second) highest score in column and section; underline = train and test set from same dataset.

Model	Seed	Hyperparameters
LLaMa 3 8B	8989	*checkpoint*=4562, *learning rate scheduler*=cosine, *learning rate*=5e-6
	79817	*checkpoint*=9124, *learning rate scheduler*=cosine, *learning rate*=5e-6
	794323	*checkpoint*=13686, *learning rate scheduler*=constant, *learning rate*=5e-6
LLaMa 3.1 8B	8989	*checkpoint*=4562, *learning rate scheduler*=cosine, *learning rate*=5e-6
	79817	*checkpoint*=13686, *learning rate scheduler*=cosine, *learning rate*=2e-4
	794323	*checkpoint*=13686, *learning rate scheduler*=constant, *learning rate*=5e-6
Mistral 7B	8989	*checkpoint*=13686, *learning rate scheduler*=constant, *learning rate*=5e-6
	79817	*checkpoint*=9124, *learning rate scheduler*=constant, *learning rate*=5e-6
	794323	*checkpoint*=4562, *learning rate scheduler*=cosine, *learning rate*=5e-6

Table 17: Hyperparameters of the best-performing modules for sentiment control, selected through grid search. The modules were trained on the combined datasets.

Model	Seed	Hyperparameters
LLaMa 3 8B	8989	*checkpoint*=27997, *learning rate scheduler*=constant, *learning rate*=5e-6
	79817	*checkpoint*=27997, *learning rate scheduler*=constant, *learning rate*=5e-6
	794323	*checkpoint*=27997, *learning rate scheduler*=constant, *learning rate*=5e-6
LLaMa 3.1 8B	8989	*checkpoint*=27997, *learning rate scheduler*=constant, *learning rate*=5e-6
	79817	*checkpoint*=13998, *learning rate scheduler*=cosine, *learning rate*=5e-6
	794323	*checkpoint*=41995, *learning rate scheduler*=constant, *learning rate*=2e-4
Mistral 7B	8989	*checkpoint*=13998, *learning rate scheduler*=cosine, *learning rate*=2e-4
	79817	*checkpoint*=13998, *learning rate scheduler*=cosine, *learning rate*=5e-6
	794323	*checkpoint*=41995, *learning rate scheduler*=constant, *learning rate*=2e-4

Table 18: Hyperparameters of the best-performing modules for topic control, selected through grid search. The modules were trained on the combined datasets.