

Parameter-Efficient Fine-Tuning of Lightweight LLMs for Persona-Based Dialogue

by

Ayan Myrzakhmet

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

May 2026

© Nazarbayev University 2026. All rights reserved.

Author
Department of Data Science
11

Certified by
Michael Lewis
Associate Professor
Thesis Coordinator

Certified by
Latif Zohaib
Assistant Professor
Thesis Coordinator

Accepted by
Elizabeth Arkhangelsky
Dean, School of Engineering and Digital Sciences

Parameter-Efficient Fine-Tuning of Lightweight LLMs for Persona-Based Dialogue

by

Ayan Myrzakhmet

Submitted to the Department of Data Science
on 11, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Language models have exhibited good performance on several natural language processing tasks. However, for persona-based generation, particularly with smaller and efficient models, there still exist difficulties. Although open-source and lightweight models may be desirable because they require fewer resources to fine-tune and deploy, they tend to struggle with adherence to the persona definition, consistency, and personality-like replies. The following research questions guide this work: Is LoRA adaptation capable of enhancing persona-based behavior in a lightweight LLM?

The study uses `microsoft/Phi-3-mini-4k-instruct` as the base model and compares it with two adapted variants: a standard LoRA adapter trained on RoleBench and a routed clustered-LoRA system trained on role-description clusters. The goal is to evaluate whether supervised LoRA-based adaptation improves the model compared with its original base version, and whether clustered adapters provide additional benefit over a single global adapter.

This evaluation includes four aspects which are RoleBench in terms of role-dependent responses, CharacterBench with regards to character customization performance, PersonaGym regarding persona-agent behavior, and customized MPI/IPPI-type test for personality-dependent responses, especially based on Big Five personality traits. Results show that the use of LoRA leads to improvements in RoleBench performance within the domain boundaries. In particular, the phi-3 model baseline showed 20.23 for RoleBench SPE score, the baseline with LoRA adaptation got 24.78, and routed clustered-LoRA achieved 26.03. However, such positive results do not necessarily apply when evaluating performance on other benchmarks. In case of CharacterBench and PersonaGym, the baseline model performed better than any adapted variants.

The evaluation via MPI/IPPI showed that by adapting the model, it was possible to adapt the personality-oriented answering styles of the modified model; the agreeableness score went up while the neuroticism score went down. In other words, the experimental findings suggest that LoRA adaptation can be useful for improving role playing performance of lightweight LLMs. Nevertheless, training on only one role-playing dataset is not enough for boosting the performance across all the

persona-oriented testing configurations. Clustered LoRA represents an appealing choice between one global adapter and fully separate persona adapters.

Keywords: large language models, lightweight LLMs, persona-based dialogue, role-playing language models, LoRA, QLoRA, parameter-efficient fine-tuning, RoleBench, CharacterBench, PersonaGym, MPI, IPIP, clustered adapters

Thesis Coordinator: Michael Lewis
Title: Associate Professor

Thesis Coordinator: Latif Zohaib
Title: Assistant Professor

Contents

1	Introduction	13
1.1	Motivation	14
1.2	Research Problem	15
1.3	Thesis Objectives	16
1.4	Contributions	17
1.5	Thesis Structure	18
2	Related Work	19
2.1	Large Language Models and Instruction-Tuned Dialogue Systems . .	19
2.2	Persona-Based Dialogue	20
2.3	Role-Playing Language Models and RoleBench	21
2.4	Character Customisation and CharacterBench	22
2.5	Persona-Agent Evaluation and PersonaGym	23
2.6	Personality-Oriented Evaluation: MPI and IPIP	23
2.7	Parameter-Efficient Fine-Tuning	24
2.8	Clustered Adapter Training	25
2.9	Research Gap	26
3	Datasets and Benchmarks	27
3.1	RoleBench	28
3.2	RoleBench Data for Clustered LoRA Training	29
3.3	CharacterBench	30
3.4	PersonaGym	31

3.5	MPI/IPIP-Style Personality Evaluation	32
3.6	Dataset Usage Summary	34
4	Methodology	35
4.1	Experimental Overview	36
4.2	Base Model Selection	37
4.3	Training Dataset and Prompt Format	37
4.4	LoRA and QLoRA Fine-Tuning	38
4.5	Standard LoRA Adapter	39
4.6	Clustered LoRA Adapter Training	39
4.7	Clustered Training Data Distribution	40
4.8	Training Environment	41
4.9	Inference Procedure	42
4.10	Evaluation Design	42
4.10.1	RoleBench Evaluation	43
4.10.2	CharacterBench Evaluation	44
4.10.3	PersonaGym Evaluation	44
4.10.4	MPI/IPIP-Style Personality Evaluation	45
4.11	Compared Configurations	45
4.12	Code Availability	46
4.13	Methodological Limitations	46
4.14	Summary	47
5	Results	49
5.1	RoleBench Results	50
5.1.1	Comparison with RoleBench Paper Results	51
5.2	CharacterBench Results	51
5.2.1	Comparison with CharacterBench Paper Results	52
5.3	PersonaGym Results	53
5.3.1	Comparison with PersonaGym Paper Results	54
5.4	MPI/IPIP-Style Personality Evaluation	55

5.4.1	Comparison with MPI Reference Results	56
5.5	Cross-Benchmark Comparison	57
5.6	Summary of Main Findings	58
5.7	Limitations of the Results	59
5.8	Summary	60
6	Discussion	61
6.1	Interpretation of the Main Results	61
6.2	Effect of Clustered LoRA	62
6.3	Personality-Oriented Behaviour	63
6.4	Implications	63
6.5	Limitations	64
6.6	Future Work	65
6.7	Closing Perspective	65
7	Conclusion	67

List of Figures

4-1	Overview of the adaptation and evaluation methodology.	36
4-2	Clustered LoRA training and routing pipeline.	40
4-3	Evaluation pipeline across RoleBench, CharacterBench, PersonaGym, and MPI/IPIP-style testing.	43

List of Tables

3.1	RoleBench splits used in this thesis	29
3.2	RoleBench training data distribution across clustered LoRA adapters	30
3.3	CharacterBench evaluation data used in this thesis	31
3.4	PersonaGym benchmark structure used in this thesis	32
3.5	Summary of dataset and benchmark usage	34
4.1	LoRA and QLoRA configuration used in the experiments	38
4.2	Training data distribution across clustered LoRA adapters	41
4.3	Main training configuration for clustered LoRA adapters	41
4.4	Model configurations compared in the thesis	46
5.1	RoleBench SPE results for Phi-3-based models	50
5.2	Selected RoleBench paper references and Phi-3 results	51
5.3	CharacterBench average scores for Phi-3-based models	52
5.4	Selected CharacterBench paper references and Phi-3 results	53
5.5	PersonaGym task-level results for Phi-3-based models	54
5.6	Selected PersonaGym paper references and Phi-3 results	55
5.7	MPI/IPIP-style OCEAN scores for Phi-3-based models	56
5.8	MPI reference results and Phi-3-based OCEAN scores	56
5.9	Summary of main results across evaluation settings	57

Chapter 1

Introduction

Large language models (LLMs) have become central fixtures in natural language processing, underpinning applications as varied as open-domain question answering, multi-step reasoning, dialogue systems, and automatic summarisation [2, 16]. The Transformer architecture introduced by Vaswani et al. [26] provided the structural backbone that made this wave of development possible, and subsequent scaling studies have confirmed that performance tends to improve predictably as models, data, and compute budgets grow [14, 23]. With wider deployment, however, a different class of problem has become harder to ignore: how to keep a deployed system consistent over time, aligned with a chosen identity, and practical to run under real-world hardware constraints.

These challenges are particularly evident in **persona-grounded dialogue**. Unlike generic text generation, persona-grounded interaction demands that the model sustain a recognisable identity across an entire conversation — a coherent set of background facts, preferences, behavioural tendencies, and, in some settings, moral or emotional boundaries. This requirement appears in customer support agents, educational tutors, social companions, role-playing platforms, and interactive game characters, all of which ask users to believe they are speaking to the same agent from one session to the next [18, 27, 29].

While researchers have made significant strides in our ability to create personas, sustaining that persona throughout a conversation is still a big challenge. Other

studies have found that LLMs tend to deviate from the role that was assigned to them; produce generic responses like those of an assistant; provide responses that contradict the attributes defined for the persona; or do not demonstrate the characteristics intended by the author of the response consistently [24, 30]. For this reason, the challenges associated with creating a consistent persona are particularly relevant for lightweight models. Models within the 1B – 4B parameter space are interesting due to the fact that they can be used at lower costs of deployment and are faster to fine tune than larger frontier models. However, smaller models also have significantly poorer performance with respect to the ability to follow instructions and act out a role [12, 25]. This leads us to an important empirical question: Can we use targeted fine-tuning to develop lightweight models that will be able to perform better on tasks related to personas?

This thesis seeks to evaluate the ability of lightweight LLMs, if exactly `microsoft’s Phi-3 mini-4k-instruct`, a lightweight instruction-tuned LLM to produce persona-based responses and determine if parameter efficient fine-tuning will increase the degree to which those models adhere to the persona [1]. Specifically, it assesses a base lightweight model compared to its LoRA fine-tuned version that was trained using both role playing data and/or data related to persona. We are not seeking to propose a novel memory architecture, but rather seek to measure what improvements can be achieved based on supervised adaptation alone.

Therefore, the primary research question addressed in this thesis is:

Can parameter-efficient fine-tuning improve the ability of lightweight language models to generate persona-consistent responses?

1.1 Motivation

Desire to conduct this research stems from two practical limitations. First, while large proprietary models typically exhibit good behavior for roles and simulations of persons, they are costly to execute, difficult to interpret, and typically cannot be fully fine tuned. Secondly, many applications utilizing models need to utilize models that

can operate with limited hardware, reduced latency, and less computational overhead. As a result, lightweight models available for public consumption represent an essential substitute to large scale frontier models.

Lightweight models do not necessarily function as reliable agents for personas. Although a small model understands the literal interpretation of “acting as” a specific role, it can still produce responses in a generic manner; ignore character-specific information; and/or fail to maintain the expected tone. Therefore, this necessitates systematic assessment. In contrast to simply assuming that a fine-tuned model performs better than a base model, the thesis determines if LoRA-based adaptation results in quantifiable improvements relative to the base model.

Low Rank Adaptation (LoRA) offers a practical means of addressing this issue [11]. LoRA adapts a pre-trained model’s weight vectors by training small rank-limited matrices added to select layers of the model. Consequently, LoRA substantially limits the number of trainable parameters resulting in fine-tuning being achievable on limited GPU resources. QLoRA further decreases memory requirements by incorporating low rank adaptation with quantized model loadings [6]. Thus, adapting lightweight models for persona-based generation is reasonable without high costs associated with full fine-tuning.

1.2 Research Problem

Several datasets and metrics have been developed in prior research for assessing persona-based dialogue, role-playing, and character customization. PersonaChat established persona-conditioned dialogue as a task in which models generate responses based upon profile statements [29]. RoleLLM and RoleBench focus on the role-playing abilities of LLMs across multiple characters [27]. PersonaGym provides an evaluation platform for persona-agents across various environments/tasks. CharacterBench measures characterization across multiple dimensions including: persona consistency; memory; knowledge boundaries; emotion; morality; and believability [24,30]. Further psychometric assessments such as MPI- and IPIP-type testing assess whether LLMs

can express stable personality-like characteristics [13, 15].

While previous research demonstrates that persona behavior can be assessed using various criteria, there is still a practical void concerning compact open source models. Most previous research focused on large models; developing benchmark datasets; proposing prompt-based techniques; and developing datasets for role-playing. There has been relatively little attention directed towards a basic yet critical deployment-related question: Will a small model be sufficiently improved via LoRA fine-tuning so as to be more dependable on standard evaluation tasks for persona-related evaluations?

This thesis attempts to bridge this void by comparing the following three items in a controlled manner:

1. a base lightweight instruction-tuned LLM;
2. the above mentioned base model after having undergone LoRA fine-tuning using on RoleBench [27];
3. clustered LoRA adapters trained on grouped RoleBench role-playing data.

The objective of this comparison is to assess if fine-tuning increases persona adherence; role-specific response quality; and stylistic coherence under standard evaluation conditions.

1.3 Thesis Objectives

This study’s primary goal was to assess the performance improvement of LoRA fine-tuning for enhancing persona-based generation with respect to lightweight LLMs. This study has the following specific goals:

- Select a lightweight open-source LLM that can be used for resource-limited fine-tuning;
- train a standard LoRA adapter on the RoleBench dataset;

- train clustered LoRA adapters using grouped role-playing data from RoleBench;
- Test both base models and fine tuned models using persona-oriented benchmarks;
- Compare these two models using both automatic metrics and rubric based assessment;
- Evaluate whether fine tuning improves overall persona consistency, role-adherence, and response-quality.

1.4 Contributions

This study will make the following contributions:

1. **Evaluation of lightweight LLMs for persona-based generation.** This study compares how well an inexpensive, open-source model performs on persona-based task(s), excluding large-scale models from consideration.
2. **LoRA-based adaptation on RoleBench.** The thesis implements parameter-efficient fine-tuning using RoleBench as the main role-playing training source.
3. **Clustered LoRA adaptation for grouped role-playing data.** In addition to a standard LoRA adapter, the thesis investigates clustered LoRA adapters trained on grouped RoleBench data.
4. **Controlled comparisons of fine-tuned vs base models.** The contribution will compare whether the addition of LoRA fine-tuning results in significant improvement of the original model.
5. **Multi-benchmark empirical evaluation.** The thesis evaluates the model variants on RoleBench, CharacterBench, PersonaGym, and a custom MPI/IPIP-style evaluation, covering role-playing ability, character consistency, persona adherence, and personality-oriented behaviour.

1.5 Thesis Structure

The remainder of the thesis is organised as follows. Chapter 2 reviews relevant prior work on LLMs, persona modelling, role-playing benchmarks, lightweight LLMs, and parameter-efficient fine-tuning. Chapter 3 describes the datasets and benchmarks used for training and evaluation. Chapter 4 presents the proposed methodology, covering model selection, LoRA fine-tuning, experimental setup, and evaluation protocol. Chapter 5 reports the experimental results comparing the base and fine-tuned models. Chapter 6 interprets those findings, discusses limitations, and situates the work within the broader field. Finally, Chapter 7 concludes and sketches directions for future research.

Chapter 2

Related Work

The purpose of this chapter is to describe previous work relating to evaluating and adapting lightweight language models to support persona-based generation. The description is structured along six lines: large language models and instruction tuning, persona-based dialogue, role-playing datasets and benchmarks, character-customisation evaluation, personality-oriented evaluation, and parameter-efficient fine-tuning. The chapter also identifies the research gap which this thesis addresses: whether LoRA-based adaptation, including clustered LoRA adapters, enhances persona-related behaviour in a lightweight LLM across multiple evaluation settings [3].

2.1 Large Language Models and Instruction-Tuned Dialogue Systems

Large language models (LLMs) are neural language models trained on large-scale text corpora to predict and generate natural language. The Transformer architecture introduced by Vaswani et al. [26] was instrumental in enabling many of the recent improvements in LLMs, allowing models to process longer sequences of tokens through self-attention mechanisms. Subsequent work demonstrated a correlation between increasing model size, data scale, and compute with higher performance [2, 10, 14]. Consequently, LLMs are now widely applied across a range of tasks including dialogue,

question answering, summarisation, reasoning, and instruction-following [3, 16, 23].

Instruction tuning represents another significant development in the practical application of LLMs. Rather than being trained solely to predict the next token, instruction-tuned models are further adapted to respond to user requests, questions, and task descriptions [20]. This improves their suitability as conversational systems, since they learn to interpret explicit prompts and act accordingly. Nevertheless, instruction tuning does not guarantee that a model will behave consistently as a particular persona or role. A model may understand an instruction such as “act as a historian” or “speak like a fictional character”, but still produce generic assistant-like responses or fail to maintain the expected role-specific behaviour.

This limitation is especially important for lightweight LLMs. Compact models are attractive because they are cheaper to fine-tune and deploy, but they often have weaker reasoning ability, instruction-following capability, and stylistic control than larger frontier models [12, 25]. Therefore, examining whether lightweight instruction-tuned models can be improved through parameter-efficient adaptation is an important practical research direction.

2.2 Persona-Based Dialogue

Persona-based dialogue aims to make conversational agents more consistent and personalised by conditioning their responses on a specific identity or profile. Early neural dialogue systems often produced generic, inconsistent, or contradictory responses because they were trained across many speakers without an explicit representation of speaker identity [18]. Li et al. [18] addressed this problem by introducing persona-based neural conversation models, where speaker embeddings were used to capture individual characteristics such as speaking style and background knowledge [17].

Zhang et al. [29] further developed the concept of persona-grounded dialogue by introducing PersonaChat. In this dataset, each participant is assigned a short multi-sentence profile describing their identity and personality, providing a method by which conversational AI can be trained to produce responses tailored to a particular persona.

Despite these advances, persona-based dialogue remains difficult. A model may correctly use persona information in one response but contradict it in another, or copy superficial words from a persona description without genuinely adapting its behaviour [24, 30]. These problems are most apparent in role-playing and character-based dialogue, where the model must preserve not only factual attributes but also tone, style, values, and role-specific knowledge.

In this thesis, persona-based dialogue is treated as a broad category encompassing profile-based personas, fictional or occupational roles, and personality-oriented behavioural descriptions. The focus is not on developing a new dialogue architecture, but on evaluating whether a lightweight LLM becomes more persona-consistent after LoRA-based fine-tuning.

2.3 Role-Playing Language Models and RoleBench

Role-playing extends persona-based dialogue by requiring a model to emulate a specific character, profession, or role. A role-playing model should not only produce correct answers but also generate responses that reflect the assigned role’s language style, background knowledge, and behavioural tendencies [17]. This is more demanding than generic instruction-following because the model must adapt its responses to a role-specific distribution.

RoleLLM, introduced by Wang et al. [27], is one of the key contributions in this area. It proposes a general framework for benchmarking, eliciting, and improving role-playing abilities in LLMs. The associated RoleBench dataset contains role-playing instruction-response examples constructed from character profiles, role-specific knowledge, and generated responses. RoleBench is especially relevant to this thesis because it provides both training data and an evaluation setting for role-playing adaptation.

RoleLLM also demonstrates that role-playing ability can be improved through role-conditioned instruction tuning, supporting the central assumption of this thesis that supervised adaptation on role-playing data may improve a model’s persona adherence. However, this thesis focuses specifically on a lightweight Phi-based model

and compares three settings: the base model, a standard LoRA-adapted model, and clustered LoRA adapters trained on grouped RoleBench data.

2.4 Character Customisation and CharacterBench

Character customisation is closely related to role-playing, but places stronger emphasis on whether a model can accurately imitate a target character across multiple dimensions. A character-customised model should effectively represent a character’s identity, factual background, behavioural style, emotional expression, and appropriate knowledge boundaries.

CharacterBench, introduced by Zhou et al. [30], provides a large-scale benchmark for evaluating character customisation in LLMs. It defines multiple evaluation aspects including memory, knowledge, persona, emotion, morality, and believability, further divided into dimensions such as memory consistency, fact accuracy, boundary consistency, attribute consistency, behaviour consistency, emotional self-regulation, empathetic responsiveness, morality stability, morality robustness, human-likeness, and engagement.

CharacterBench is useful because it evaluates character behaviour more comprehensively than simple role-playing accuracy. A model may produce fluent and stylistically rich responses while still violating a character’s knowledge boundary or failing to preserve character attributes, making CharacterBench well-suited for testing whether LoRA adaptation improves deeper character consistency rather than only surface-level response quality.

In this thesis, CharacterBench is used as an external evaluation benchmark. The model is not trained on CharacterBench data; rather, it is used to assess whether improvements gained through RoleBench adaptation generalise to a different character-evaluation setting. This distinction is important because it allows the thesis to evaluate whether LoRA and clustered LoRA adapters improve general persona-related behaviour beyond the original training dataset.

2.5 Persona-Agent Evaluation and PersonaGym

PersonaGym provides an alternative perspective on persona evaluation [24]. Rather than focusing solely on dialogue profile consistency or fictional-character imitation, PersonaGym evaluates persona agents across environments and tasks, and introduces PersonaScore, an automatic metric designed to assess how well a persona agent behaves according to its assigned persona.

PersonaGym evaluates persona agents through several task types grounded in decision-making and persona-relevant behaviour, including expected action, linguistic habits, persona consistency, toxicity control, and action justification. This makes PersonaGym useful for evaluating whether a model can act as a persona agent across many circumstances beyond generating stylistically similar responses.

This benchmark is especially relevant to this thesis because it helps evaluate persona behaviour from a broader agent-oriented perspective. A model that performs well on RoleBench may have learned role-specific response patterns, but PersonaGym can test whether that adaptation also improves behaviour across different persona-relevant environments and question types. Therefore, PersonaGym is used as one of the main evaluation settings for comparing the base model, standard LoRA adapter, and clustered LoRA adapters.

2.6 Personality-Oriented Evaluation: MPI and IPIP

Persona and personality are related but not identical concepts. A persona typically describes a role, character, profession, or background profile, while personality refers to more stable behavioural tendencies, commonly described through psychological trait models. The Big Five model, also known as OCEAN, defines personality through five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [5].

Jiang et al. [13] introduced the Machine Personality Inventory (MPI) as a method for evaluating personality-like behaviour in language models. MPI adapts psychometric-

style testing to LLMs by asking models to respond to questionnaire items associated with Big Five traits, enabling researchers to obtain structured personality scores and assess whether prompting or adaptation changes a model’s expressed behavioural tendencies.

Similarly, IPIP-based evaluations use items derived from the International Personality Item Pool. Klinkert et al. [15] used IPIP-based personality assessments to compare LLM outputs with human personality profiles, demonstrating that models can exhibit considerable variability when attempting to simulate human-like personality patterns.

In this thesis, MPI/IPIP-style testing is used as a custom evaluation component. The purpose is not to claim that the model has a real personality, but to investigate whether its responses to personality-oriented prompts change after LoRA adaptation. This provides a complementary perspective to RoleBench, CharacterBench, and PersonaGym: while those benchmarks evaluate role-playing and persona adherence, MPI/IPIP-style testing examines whether adaptation affects broader personality-like response patterns.

2.7 Parameter-Efficient Fine-Tuning

Fine-tuning all parameters of an LLM is computationally expensive, particularly for users with limited GPU resources. Parameter-efficient fine-tuning methods address this by updating only a small subset of parameters while keeping most of the base model frozen, making them especially useful for adapting lightweight models to specialised tasks [4].

Low-Rank Adaptation (LoRA), introduced by Hu et al. [11], is one of the most widely used parameter-efficient fine-tuning methods. LoRA freezes the original model weights and inserts trainable low-rank matrices into selected layers, greatly reducing the number of trainable parameters while still enabling the model to adapt to new tasks or styles.

QLoRA extends this approach by integrating quantised model loading with low-

rank adaptation [6]. By training LoRA adapters on top of a base model loaded in low precision, QLoRA further reduces memory consumption and makes fine-tuning feasible on more limited hardware, which is especially relevant for lightweight LLM experimentation without requiring large-scale compute infrastructure.

For persona-based generation, LoRA is attractive because it enables adapting a model toward role-specific response patterns without modifying the underlying backbone. This allows the same base model to be compared directly before and after adaptation, and also enables multiple adapters to be trained for different subsets of roles or persona types.

2.8 Clustered Adapter Training

A standard LoRA setup typically trains a single adapter over the entire training dataset. While straightforward, this approach assumes that all examples should influence the model equally. In role-playing data, however, different roles may require different linguistic styles, knowledge patterns, and behavioural tendencies. For example, a scientific expert, a historical figure, and a fictional fantasy character may each require different response distributions.

Clustered adapter training addresses this by first dividing role-playing data into clusters and then training a separate adapter for each group. The motivation is that each adapter can specialise in a narrower subset of role or persona behaviour while still relying on the same frozen base model, potentially reducing interference between substantially different roles and enabling more targeted adaptation. This clustered adapter configuration connects to past studies concerning modular adapter architectures and combinations of LoRA experts, in which multiple adapters serve as dedicated parts rather than training a single adapter for all inputs [21, 22, 28].

In this thesis, clustered LoRA adapters are compared with a standard LoRA adapter. The intent is not to propose a novel model architecture, but to test whether grouping role-playing data before adapter training produces measurable differences. This comparison is important because clustered adapters may offer a practical middle

ground between a single general adapter and a separate adapter for every individual persona.

2.9 Research Gap

The reviewed literature shows that persona-based dialogue, role-playing, character customisation, and personality-oriented evaluation have each developed into active research areas. PersonaChat introduced profile-grounded dialogue; RoleBench supported role-playing instruction tuning; CharacterBench evaluated character customisation across multiple dimensions; PersonaGym evaluated persona agents across persona-relevant tasks; and MPI/IPIP-style methods provided a structured framework for personality-oriented evaluation.

However, several gaps remain. First, much prior work focuses on large models or benchmark construction, while fewer studies examine compact open-source models under resource-constrained adaptation. Second, many studies evaluate one benchmark at a time, making it difficult to determine whether improvements on one persona dataset transfer to other persona-oriented evaluation settings. Third, standard LoRA adaptation is often treated as a single-adapter process, leaving the potential value of clustered adapters for role-playing data underexplored.

This thesis addresses these gaps by evaluating `microsoft/Phi-3-mini-4k-instruct` as a lightweight base model, adapting it with both standard LoRA and clustered LoRA adapters using RoleBench, and evaluating the resulting models across RoleBench, CharacterBench, PersonaGym, and a custom MPI/IPIP-style evaluation. This design allows the thesis to examine not only whether LoRA adaptation improves persona-based behaviour, but also whether those improvements are consistent across different types of persona evaluation.

Chapter 3

Datasets and Benchmarks

This chapter introduces the datasets and benchmarks employed in this research. One major training resource with multiple supplemental evaluation resources constitute the design of the experiments in this research. RoleBench is utilized as the principal dataset for LoRA-based adaptation, including the conventional adapter along with the clustering adapter configuration. Additionally, RoleBench has been used for role-play based evaluation. CharacterBench, PersonaGym, and the custom MPI/ IPIP-Style Evaluation have been employed as supplementary external testing environments for assessing if adaptation impacts from LoRA can be transferred across different sources of training data. The organization of this chapter is as follows. Section 3.1 discusses the use of RoleBench in training and evaluating LoRA adaptations. Section 3.2 outlines how RoleBench was segmented to support clustered LoRA Training. Section 3.3 illustrates CharacterBench as an external benchmark for character customization. Section 3.4 outlines the use of PersonaGym as a framework for agent-persona based evaluation. Finally, Section 3.5 presents the psychometric MPI/ IPIP-Style Evaluation. Section 3.6 provides an overview of how all of these datasets contribute to the overall objectives of this research.

3.1 RoleBench

RoleBench is the main dataset employed throughout this thesis. It was developed within the context of RoleLLM (a platform for benchmarking, acquiring, and improving role-playing capabilities of language models) [27]. RoleBench contains role-conditioned instructional response pairs intended to instruct and measure role-playing behavior in language models. Unlike other simple datasets such as persona-profile dialogue datasets, RoleBench includes role specific instructions, character oriented responses and examples that illustrate both role knowledge and communication style. RoleBench has been used twice in this thesis. Firstly, it served as the training source for LoRA adaptation; both the traditional LoRA Adapter and Clustered LoRA Adapters were trained using RoleBench. Secondly, RoleBench has been employed as an evaluation benchmark to determine differences in performance between the base model and adapted models on role-specific response generation.

The dataset repository used in the experiments is:

`ZenMoore/RoleBench`

The main training file used for both the global LoRA adapter and clustered LoRA adapters is:

`rolebench-eng/role-generalization/general/train.jsonl`

This split contains 118,035 training examples. Each training example is converted into an instruction following format compatible with Phi-3 chat template. At a high-level, each example included information about a particular role or person (the 'role'), a user query or request ('instruction') that needed to be fulfilled by the role and the 'expected' assistant response based upon that role's personality or perspective.

RoleBench originally reported four separate categories of evaluation metrics. These include Instructional Behavior (SPE) and Role-Specific Behavior (RAW,CUS,SPE). However, in this research, the saved RoleBench Outputs recovered only the SPE Component (which measures Rouge-L Evaluation Metrics on role-specific instructional responses). The present thesis is mainly concerned with assessing the retrieved

Table 3.1: RoleBench splits used in this thesis

Split / file	Size	Use in thesis
role-generalization/ general/train.jsonl	118,035	Training source for standard LoRA and clustered LoRA adapters.
role-generalization/ general/test.jsonl	5,000	Role-generalization test split; used in the evaluation setup, although general examples were not successfully aligned in the saved RoleBench outputs.
role-generalization/ role_specific/test.jsonl	2,534	Role-specific evaluation split used for routed clustered-adaptor RoleBench evaluation.
instruction-generalization	partially scored	Used for partial base-model and global-adaptor paper-style evaluation; saved outputs recovered only role-specific SPE scores.

SPE component that measures Rouge-L against role-based instructions. It should be noted that the full RAW/CUS/SPE replication was not part of the assessment scenario.

3.2 RoleBench Data for Clustered LoRA Training

To train a single LoRA Adapter based on the whole RoleBench training split and in addition to train clustered LoRA Adapters. Training a clustered LoRA adapter has its basis in the assumption that role playing data contains a lot of different types of characters and roles and that a single adapter will have to mix many different types of writing styles together. With Clustered adapters you divide your training data into several small groups (clusters) and train one adapter for each group. For this experiment five clusters were chosen. Thus the total number of examples (118,035) of the training data from RoleBench were distributed over the five clusters by their role assignments. All five clusters together represent 80 different types of roles. The clustering procedure itself was described in Chapter 4; the purpose of this chapter is to report the results of the clustering process (the division of the training data).

However, there is a drawback of the above mentioned way of dividing the training

Table 3.2: RoleBench training data distribution across clustered LoRA adapters

Cluster	Roles	Training examples
cluster_0	36	53,106
cluster_1	7	10,322
cluster_2	13	19,136
cluster_3	14	20,743
cluster_4	10	14,728
Total	80	118,035

data into smaller groups. Due to the filter of the role-generalization test split regarding cluster role membership, all five clusters were empty when trying to calculate cluster specific validation sets. Therefore, all five clustered adapters had to be trained without in-training validation metrics. While this doesn’t make it impossible to train the adapters; it makes sense to use a validation metric while training the adapters. After they were trained, the trained adapters were evaluated again using RoleBench, CharacterBench, PersonaGym, and the custom MPI/IPIP style evaluation.

3.3 CharacterBench

The character customization quality is evaluated by using CharacterBench as an external benchmark. CharacterBench is especially suited because it evaluates role-playing systems across multiple dimensions instead of only evaluating the general response quality. CharacterBench includes dimensions like memory, knowledge, persona, emotion, morality, and believability [30]. Additionally, these aspects are further specified through the following sub-dimensions like memory consistency, fact accuracy, boundary consistency, attribute consistency, behavior consistency, emotional self-regulation, empathetic responsiveness, morality stability, morality robustness, human-likeness, and engagement. In this thesis CharacterBench wasn’t used as part of the training process. Instead, it was used to check whether models which were adapted on RoleBench show some level of transfer to a different character-evaluation-benchmark. This difference is essential: If a LoRA Adapter improves only on RoleBench but not on CharacterBench, then we would say that the LoRA Adapter adapts only to a dataset and is therefore not universally applicable for personas or

Table 3.3: CharacterBench evaluation data used in this thesis

Item	Value
Total CharacterBench samples	22,859
Characters	3,956
Character categories	25
In-domain test samples	1,625
Out-of-domain test samples	1,625
Total test samples used for final evaluation	3,250
Local evaluation files	13
Samples per local evaluation file	250

behaviors of characters. The full version of CharacterBench provides 22,859 human annotated example sentences to characterize 3,956 characters from 25 character categories [30]. For evaluation purposes the test set provided by CharacterBench is divided into two parts: an in-domain subset and an out-of-domain subset – both having 1,625 samples respectively. So together there are 3,250 test samples. In the local benchmark files which are used within this thesis these 3,250 samples are organized into 13 files for evaluation purposes with each file containing 250 sessions.

The characterization of the performance of the models is done by producing model answers for the test samples and afterwards rating them using CharacterJudge – the judge-model which corresponds to CharacterBench. Since the testset is identical for all models in order to enable a fair comparison of the base model, standard LORa Adapter, cluster-specific Adapter and routed-clustered Adapter-system.

3.4 PersonaGym

PersonaGym was selected to serve as another assessment tool for evaluating a persona agent in addition to RoleBench and CharacterBench [24]. The key difference between these tools and PersonaGym lies in the fact that PersonaGym assesses how well a persona agent functions within a variety of different environment/task combinations that relate to personas, whereas RoleBench and CharacterBench primarily evaluate instructional response pairs and characterization dialogue aspects. In order to facilitate this, PersonaGym introduced PersonaScore; an automated metric that measures how effectively a model performs with respect to a particular persona [24]. The Per-

Table 3.4: PersonaGym benchmark structure used in this thesis

Item	Value
Personas in original paper	200
Questions in original paper	10,000
Personas in local benchmark folder	198
Tasks per persona	5
Questions per task	10
Questions per persona	50
Total local prompts	9,900
Evaluation metric	PersonaScore

sonaGym benchmark assesses the performance of persona agents in the context of various categories, such as expected actions, language usage, persona consistency, controlling toxicity and providing explanations for their behavior. These categories provide value in that they help determine if a model can function as a persona in different contexts beyond simply mimicking the surface level talking characteristics of a persona. Within this thesis, the use of PersonaGym is limited to the assessment of the models’ performance. Therefore, no PersonaGym data will be used during training of the LoRA adapters. This allows PersonaGym to act as an independent assessment of whether adaptations made through RoleBench increase the overall ability of a persona agent to function across varying personas. The same model variations were assessed via PersonaGym as those previously tested on RoleBench and CharacterBench. The three model variants included were:

- the base `microsoft/Phi-3-mini-4k-instruct` model;
- the standard RoleBench LoRA adapter;
- and the cluster-based LoRA adapter.

3.5 MPI/IPIP-Style Personality Evaluation

The final assessment area was a custom MPI/IPIP-type personality evaluation. The basis for this assessment is that language models can be queried using psychometric-type questionnaires. The International Personality Item Pool (IPIP) provides open-

source questionnaire items related to the Big Five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [8, 9]. The Machine Personality Inventory (MPI), however, extends this concept to include the development of questionnaire items about personalities in a model-friendlier multiple choice format [13].

The MPI/IPIP assessment was developed as an automated multiple choice probe in this thesis. Each item contained a behavioral description associated with a specific Big Five personality trait along with a polarity indicator. The polarity indicator indicated whether agreement with the item would result in either an increase or decrease in the corresponding Big Five personality score. The model was presented with each item and requested to respond with one of five Likert scale options which ranged from “very accurate” to “very inaccurate”.

Using the scoring protocol for MPI, positive keyed items were scored from 5 to 1, and negative keyed items were reverse scored. A mean score was then calculated for each Big Five personality dimension. These scores were used to compute three metrics: target trait alignment, trait controllability, and test-retest consistency.

This type of assessment was not intended to conclude that the model possessed a real psychological personality. Instead, the scores were utilized as standardized behavioral probes to compare whether both the base model and LoRA-adapted models followed Big Five-type personality trait descriptions. Similar to assessing RoleBench, CharacterBench and PersonaGym, MPI/IPIP evaluations utilize trait-oriented response patterns.

This evaluation process was entirely automated and did not require human annotations. As such, it provided a repeatable means to compare the base model against the standard LoRA adapter, and cluster-based LoRA adapter(s) under identical prompts and scoring conditions.

Additional persona-dialogue datasets, such as PersonaChat and ConvAI2, are covered in Chapter 2 due to their importance in developing background knowledge regarding research concerning persona-based dialogue. However, none of these datasets were included as part of the final experimental pipeline. Experiments conducted in

Table 3.5: Summary of dataset and benchmark usage

Dataset / benchmark	Training or evaluation	Use in thesis
RoleBench	Training and evaluation	Main source for standard LoRA and clustered LoRA training; also used for role-playing evaluation.
CharacterBench	Evaluation	External character-customisation evaluation on the full 3,250-sample test set.
PersonaGym	Evaluation	External persona-agent evaluation across persona-oriented tasks.
MPI/IPIP-style evaluation	Evaluation	Automatic Big Five-style multiple-choice probing used to measure trait alignment, controllability, and test-retest consistency.

this thesis focused upon utilizing RoleBench for adapting and utilizing RoleBench, CharacterBench, PersonaGym, and MPI/IPIP-style assessments for evaluating.

3.6 Dataset Usage Summary

Overall, the structure of the dataset is reflective of the primary objective of this thesis: namely determining whether LoRA-based adaptation leads to improved behavior relevant to persons for a light-weight model. Adaptation occurs using the supervised role-playing data provided by RoleBench while CharacterBench, PersonaGym and MPI/IPIP-style assessments are used to determine whether improvements to behavior generalize across differing assessment viewpoints.

Chapter 4

Methodology

In this chapter, the methodology used for testing parameter-efficient adaptation on persona-oriented behaviour of a lightweight language model is presented. For the experiment, a controlled methodology has been chosen. This means that both a baseline and adapted versions of the base model will be tested on the same tasks. The baseline will be compared against a global LoRA adapter trained on RoleBench as well as cluster-specific LoRA adapters trained on subsets of RoleBench containing corresponding role categories.

The purpose of this particular methodology is to address two research questions at once. In the first place, this methodology is needed to see whether LoRA fine-tuning makes the persona, role-playing and character-consistent behaviour of the base model better than it was in the baseline version. Secondly, this methodology aims to see whether clustering of roles that is trained on groups of similar roles helps obtain even better results.

The thesis focuses on supervised parameter-efficient adaptation and evaluates whether this adaptation alone produces measurable improvements across several persona-oriented benchmarks.

4.1 Experimental Overview

The experimental pipeline consists of four main stages: selecting a lightweight instruction-tuned base model; preparing RoleBench examples for supervised LoRA fine-tuning; training standard and clustered LoRA adapters; evaluating the resulting model variants on persona-oriented benchmarks.

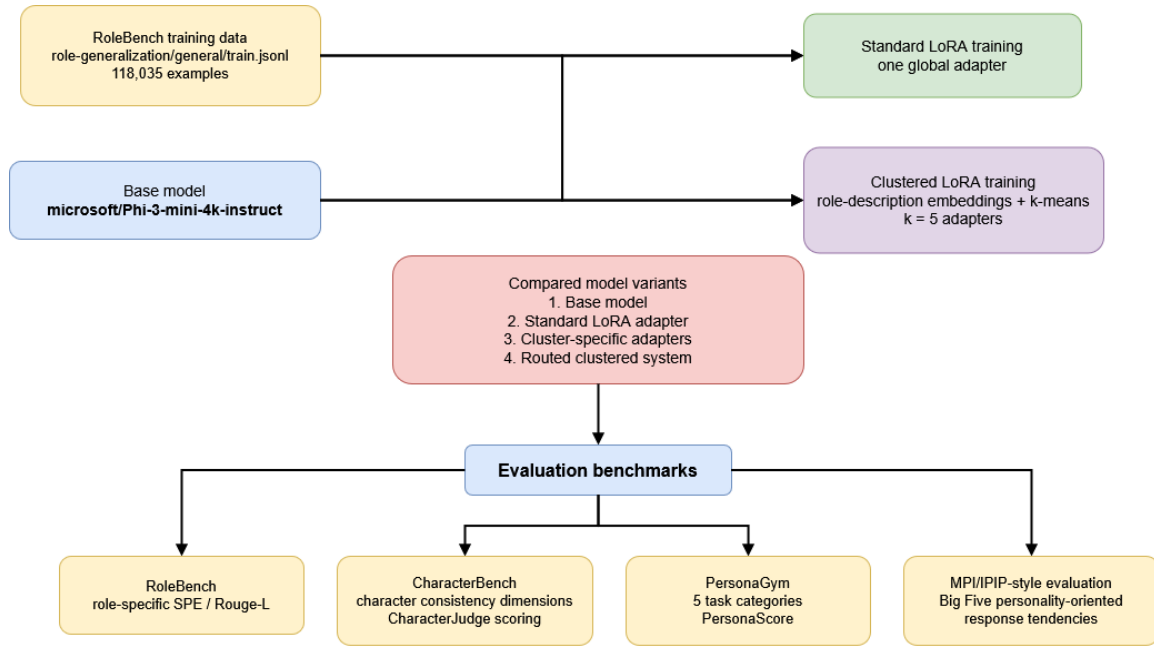


Figure 4-1: Overview of the adaptation and evaluation methodology.

The compared model variants are:

- **Base model:** the original microsoft/Phi-3-mini-4k-instruct model without additional fine-tuning.
- **Standard LoRA adapter:** a single adapter trained on the full RoleBench training split used in this thesis.
- **Clustered LoRA adapters:** five adapters trained separately on role clusters derived from RoleBench role descriptions.
- **Routed clustered system:** an inference configuration in which each input persona, role, or character is assigned to the closest role cluster and generated using the corresponding adapter.

This design allows the thesis to compare the original model, a general role-playing adapter, and a more specialised clustered-adapter setup under the same evaluation framework.

4.2 Base Model Selection

The base model used throughout the thesis is `microsoft/Phi-3-mini-4k-instruct`.

The chosen model was chosen due to its compact size and instructional tuning ability [1]. This model’s compactness allows for efficient fine-tuning via LoRA and QLoRA techniques, whereas its instructional tuning allows for using it in persona and role-play prompting.

Using a lightweight model is one of the core points of this thesis. It is important not to prove how effective a frontier model would perform in persona generation tasks but to evaluate the efficiency of parameter-efficient tuning of an open-source lightweight model.

The 4k context version was chosen because of providing enough context length for the benchmarks used in this work and being easy on computational resources.

4.3 Training Dataset and Prompt Format

RoleBench is used as the main training source for adapter fine-tuning. The training split used in this thesis is `rolebench-eng/role-generalization/general/train.jsonl`.

This split contains 118,035 training examples. Each example contains a role, a user instruction or question, and a target assistant response. The examples are converted into a supervised instruction-following format compatible with the Phi-3 chat template.

The training format follows the structure:

System: You are role-playing as {role}. Stay in character.

User: {question}

Assistant: {answer}

Table 4.1: LoRA and QLoRA configuration used in the experiments

Parameter	Value
Base model	microsoft/Phi-3-mini-4k-instruct
Quantization	4-bit NF4
Compute dtype	bf16
LoRA rank r	16
LoRA alpha	32
LoRA dropout	0.05
Optimizer	paged_adamw_8bit
Learning rate	2×10^{-4}
Epochs	1
Packing	false
Group by length	true
Seed	42

The system prompt clearly identifies the role to be played. The user prompt includes the RoleBench instruction or query. Finally, the assistant prompt includes the output response for the given instruction/role. In this manner, the system learns to generate outputs in line with the role provided.

The problem definition here is straightforward causal language modelling. With the prompt and response provided in a serialized form, the task involves predicting the next token. For LoRA fine-tuning, the base model remains unchanged, while adapter weights are updated.

4.4 LoRA and QLoRA Fine-Tuning

The thesis uses Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning [11]. LoRA freezes the original model weights and introduces trainable low-rank matrices into selected layers. This allows the model to adapt to a new task while updating only a small number of additional parameters.

The experiments use QLoRA-style loading, where the base model is loaded with 4-bit quantization and LoRA parameters are trained on top of the quantized backbone [6]. This reduces memory usage and makes fine-tuning feasible on limited GPU resources.

LoRA and QLoRA configuration used in the experiments:

The LoRA adapters are applied to both attention and feed-forward projection layers. The target modules are: `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`.

These modules allow the adapters to influence both attention behaviour and the transformation layers involved in generation.

4.5 Standard LoRA Adapter

The standard LoRA adapter is trained on the full RoleBench training split used in this thesis. It is a single global adapter shared across all roles in the training data. This configuration tests whether general role-playing adaptation improves the base model’s behaviour.

The standard adapter serves as the main fine-tuned comparison against the base model. If this adapter improves over the base model, then RoleBench LoRA adaptation is beneficial in general. If it does not improve, then supervised adaptation may either be insufficient, too narrow, or not aligned with the evaluation benchmarks.

4.6 Clustered LoRA Adapter Training

Apart from the general LoRA adapter, we train clustered LoRA adapters. Our rationale is that RoleBench has heterogeneous roles characterized by diverse linguistic characteristics, behavior patterns, and domain-specific knowledge. While a single global adapter may average out these patterns, clustering would enable specialized adapters to be trained for more homogeneous clusters.

Our pipeline for training clustered adapters comprises four steps: Extracting role descriptions from the RoleBench training set; Embedding each role description with a sentence-level embedding model; Clustering the role embeddings with k-means clustering; Training a LoRA adapter for each role cluster.

The role descriptions are embedded using `sentence-transformers/all-MiniLM-L6-v2`.

The embeddings are normalized and grouped with k-means clustering. The fi-

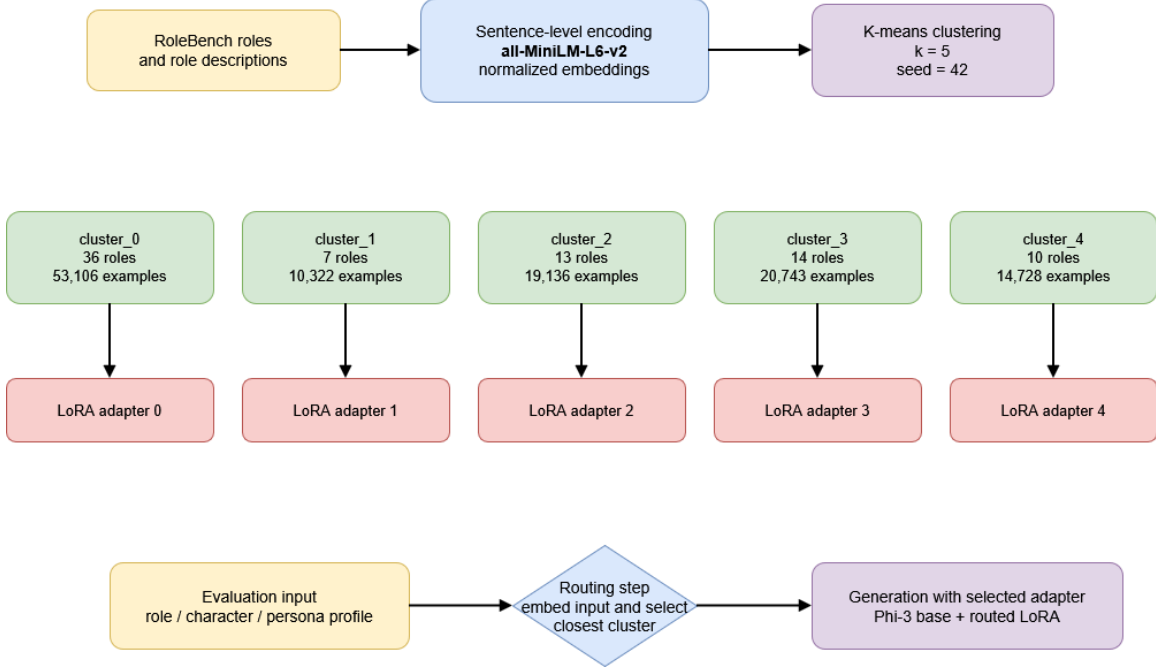


Figure 4-2: Clustered LoRA training and routing pipeline.

nal clustered setup uses: number of clusters $k = 5$; clustering algorithm: k-means; random seed: 42.

Following the clustering step, each RoleBench training example is allocated to the role cluster it belongs to. The next step involves training a different LoRA adapter for each cluster based on the examples assigned to it. This results in five distinct adapters based on clusters.

Conceptually, the clustered configuration modifies the adaptation component from having a generic adapter for all personas to having an adapter for each group of personas. During the inference phase, an input persona, role, or character is matched to the most similar role cluster, and the adapter is selected for generating text.

4.7 Clustered Training Data Distribution

The clustered setup divides the 118,035 RoleBench training examples across five clusters. The resulting distribution is:

The imbalance between clusters reflects the distribution of roles and examples

Table 4.2: Training data distribution across clustered LoRA adapters

Cluster	Roles	Training examples
cluster_0	36 roles	53,106 training examples
cluster_1	7 roles	10,322 training examples
cluster_2	13 roles	19,136 training examples
cluster_3	14 roles	20,743 training examples
cluster_4	10 roles	14,728 training examples
Total	80 roles	118,035 training examples

Table 4.3: Main training configuration for clustered LoRA adapters

Parameter	Value
Per-device batch size	4
Gradient accumulation steps	4
Effective batch size	16
Epochs	1
Learning rate	2×10^{-4}
Dataloader workers	4
Logging steps	25
Save steps	1000
Precision	bf16
Quantization	4-bit NF4
Attention implementation	eager

in the RoleBench training split. This imbalance is important because larger clusters receive more update steps and may learn more stable adapter behaviour, while smaller clusters may be more sensitive to data sparsity.

4.8 Training Environment

The adapter training was performed in GPU environments suitable for QLoRA fine-tuning. The clustered adapters were trained using an NVIDIA RTX PRO 4500 GPU. The standard full RoleBench adapter was trained separately in a larger university cluster environment.

Main training configuration for clustered LoRA adapters:

The effective batch size is computed as: effective batch size = per-device batch size $\times$ gradient accumulation steps. Thus, the clustered adapter training uses an effective batch size of $4 \times 4 = 16$.

Cluster-specific in-training validation was not used because the cluster-level held-

out splits did not provide usable examples for every cluster. Therefore, final comparison is based on post-training evaluation across the benchmarks described later in this chapter.

4.9 Inference Procedure

The inference procedure depends on the model variant being evaluated.

In terms of the base model, prompts are simply passed directly to the `microsoft/Phi-3-mini-4k-instruct` model without an adapter being loaded. For the baseline LoRA approach, the global RoleBench adapter is loaded onto the above-mentioned base model. For the clustering approach, a chosen one of the five cluster adapters trained during clustering is used.

In terms of the routing clustered method, the role, character, or persona information is used to identify the closest adapter. However, there are differences in the routing process depending on the benchmark used – RoleBench uses the associated role from the evaluation example, whereas CharacterBench character data is mapped to the nearest role cluster by means of a character’s name/profile, PersonaGym persona descriptions to the closest role-description cluster, and for the MPI/IPIP benchmarks the same models under the specified prompt format are evaluated but no role-based routing unless a persona condition is provided.

This approach investigates the adaptability of cluster adapters trained on RoleBench data for different evaluation settings.

4.10 Evaluation Design

Evaluations were conducted to compare the baseline, standard LoRA adapter, and clustered LoRA models in four different scenarios: RoleBench; CharacterBench; PersonaGym; and personality evaluation using MPI/IPIP.

Each scenario evaluates a distinct aspect of persona-related behavior. RoleBench evaluates role-based instruction responses, while CharacterBench measures charac-

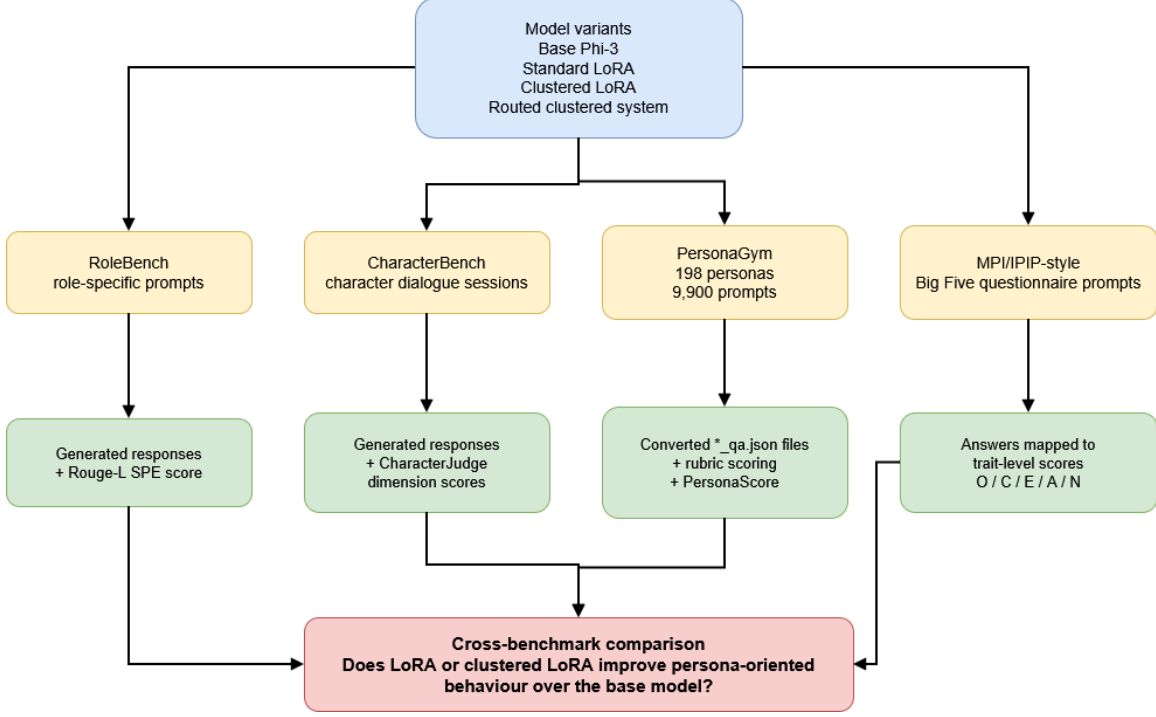


Figure 4-3: Evaluation pipeline across RoleBench, CharacterBench, PersonaGym, and MPI/IPIP-style testing.

ter customization performance in multiple aspects. PersonaGym tests persona-agent behavior for various tasks. The MPI/IPIP style test evaluates personality-related response behaviors.

4.10.1 RoleBench Evaluation

RoleBench is employed for assessing role-specific generation capabilities. The principal RoleBench evaluation criterion discussed in this thesis is SPE, calculated according to Rouge-L for role-specific instructions answers. It evaluates lexical similarity between the generated instruction response and the role-specific instruction reference answer.

The RoleBench evaluation technique is applied to the base model, regular LoRA adapter, and the routed clustered-adapters architecture. As RoleBench is the source of training data as well, this evaluation mainly reflects in-domain role-playing capabilities.

4.10.2 CharacterBench Evaluation

CharacterBench is employed as an external benchmark for character customisation [30]. In contrast with RoleBench, CharacterBench was not employed during training. Therefore, CharacterBench presents an appropriate environment to assess transferability of the role adaptation to an alternative character evaluation setup.

CharacterBench evaluates multiple aspects of character conduct, such as: memory coherence; factuality; boundary coherence; attribute coherence; behavior coherence; emotional regulation; empathy; morality coherence; morality robustness; humanness; engagement.

First, the model responses to the prompts from CharacterBench are generated. Then, the response scores are evaluated by CharacterJudge, which is the associated judging model. For all model variants, the same evaluation dataset is utilised.

4.10.3 PersonaGym Evaluation

PersonaGym is used to evaluate persona-agent behaviour across task categories and persona-relevant situations [24]. The local benchmark version used in this thesis contains 198 personas and 9,900 prompts.

PersonaGym evaluates five categories: Expected Action; Toxicity; Linguistic Habits; Persona Consistency; Action Justification.

The PersonaGym workflow consists of three stages: generate one model response per PersonaGym prompt; convert the generated responses into PersonaGym-compatible `*_qa.json` files; run the PersonaGym evaluator to obtain task-level scores and an overall PersonaScore.

The scoring hierarchy is: question-answer pairs $\rightarrow$ batch scores $\rightarrow$ task scores $\rightarrow$ PersonaScore.

PersonaGym is used only for evaluation. No PersonaGym examples are used for adapter training.

4.10.4 MPI/IPIP-Style Personality Evaluation

The custom MPI/IPIP-style evaluation is used to compare personality-oriented response tendencies before and after adaptation. This evaluation is inspired by the Machine Personality Inventory and IPIP-style Big Five questionnaires [13, 15].

The evaluation focuses on the Big Five traits: Openness; Conscientiousness; Extraversion; Agreeableness; Neuroticism.

The focus of the evaluation here is not to test whether the model has a real personality or not; rather, the purpose is to assess whether the model is capable of responding in a consistent manner like having a personality and whether such responses differ when the model is fine-tuned using RoleBench’s LoRA. All models respond to standardized questionnaire questions whose answers are analyzed based on traits.

In the case of the routed clustered-LoRA architecture evaluated through MPI/IPIP, we opted for a static clustered-adapter model rather than an adaptive routing one per item. This is due to the nature of MPI/IPIP items as general personality questionnaires without any metadata associated with the item role/persona that could align with RoleBench clusters. Consequently, the results presented for the OCEAN scores of the routed clustered-LoRA model are based on the average behavior of the static adapters for the same questions.

4.11 Compared Configurations

Model configurations compared in the thesis:

The comparison is intended to determine whether LoRA adaptation improves persona-oriented behaviour and whether clustered specialization provides additional benefit.

Table 4.4: Model configurations compared in the thesis

Configuration		Description
Base model		Original <code>microsoft/Phi-3-mini-4k-instruct</code> without additional fine-tuning.
Standard LoRA adapter		Single global adapter trained on the full RoleBench training split used in this thesis.
Cluster-specific adapter		Individual LoRA adapter trained on one role-description cluster.
Routed adapters	clustered	Five clustered adapters with routing based on role, character, or persona description.

4.12 Code Availability

The implementation code is available in the project repository: [GitHub repository for this thesis](#).

The repository contains scripts for RoleBench based training, clustered adapter construction, benchmark generation, and evaluation on RoleBench, CharacterBench, PersonaGym, and MPI/IPIP-style prompts and all the trained adapters.

4.13 Methodological Limitations

Several methodological limitations should be noted.

First, the effectiveness of clustered adapters is determined by how successfully the embeddings describe their corresponding personas and by the manner in which clusters are assigned. In cases where a test persona does not align well with any RoleBench clusters, routing will result in an adapter that is not very similar to the targeted persona.

Second, the training data for clustered adapters are very unbalanced; some clusters receive many more examples than other clusters, and this may impact the stability and generalizability of the generated adapters.

Third, various benchmarks for evaluation such as RoleBench, CharacterBench, PersonaGym, and MPI/IPIP tests test very different aspects of persona behaviors. An advancement in performance on one test may not occur on other benchmarks. As such, performance results are given from multiple testing benchmarks.

Finally, while LoRA adaptation changes the model’s tendency to generate certain responses, it does not create a memory system for retrieval or storage. Thus, any improvement in persona consistency must come from supervised training.

4.14 Summary

This chapter describes the methodology used for the thesis research. The starting point model for fine-tuning is `microsoft/Phi-3-mini-4k-instruct`. Fine-tuning of the model is achieved using LoRA/QLoRA. Training datasets are sourced from RoleBench. Two different methods of fine-tuning were examined: the use of one global LoRA adapter and five clustered LoRA adapters trained on clusters based on role-description clusters. These models are then tested against RoleBench, CharacterBench, PersonaGym, and a new personality test based on MPI/IPIP-style questions. This method is used to answer the key research questions posed by the thesis: (1) Does parameter-efficient fine-tuning improve persona-based behavior of a light-weight LLM?, and (2) Does the use of clustered adapters offer any advantages over a non-clustered adapter?

Chapter 5

Results

This chapter discusses the experimental results that evaluate how adapting LoRa via RoleBench improves the persona-oriented behavior of a lightweight language model. The base model selected was Microsoft’s `Phi-3-mini-4k-instruct`. In addition to the original base model, there were two other model types that were used for the experiments: (1) A standard LoRA adapter that was trained using RoleBench; and (2) A routed clustered LoRA system, that consisted of five LoRA role cluster adapters.

There are four groups of benchmarks used in this evaluation:

- RoleBench – provides an assessment of role specific response generation by the model;
- CharacterBench – assesses the quality of character customizations provided by the model;
- PersonaGym – provides an assessment of persona agent behavior over multiple task categories;
- MPI/IPIP style assessment – provides a measure of the Big Five personality oriented tendency of the responses produced by the model.

All the benchmark papers reported results for various models, evaluation environments, prompts and judge models. As such, they were referenced here as contextually

Table 5.1: RoleBench SPE results for Phi-3-based models

Model / system	Size / type	SPE $\uparrow$
Phi-3 base model	3.8B	20.23
Phi-3 + standard LoRA	3.8B + LoRA	24.78
Phi-3 + routed clustered LoRA	3.8B + 5 LoRA adapters	26.03

relevant reference point rather than as strict baseline controls. It should be noted that the intent here is not to establish comparative equivalence to all systems reported in the literature, but instead to provide insight into where these lightweight Phi-3-based models exist relative to much larger models, closed source models, and models specifically adapted to each benchmark.

5.1 RoleBench Results

RoleBench is the primary in-domain benchmark since it is also the source of the supervising learning data for the LoRA adaptation process. The SPE component is the focus of this evaluation since it quantifies role-specific instruction responses generated by the model using Rouge-L. Higher SPE values indicate greater lexical similarity between the response generated by the model and the reference role-specific response.

The RoleBench paper reports the results from three major Rouge-L components: RAW, CUS and SPE [27]. RAW refers to the evaluation of responses generated by the model when responding to regular instruction-following questions without regard to role specificity. CUS refers to the evaluation of customized responses to general instructions. SPE refers to the evaluation of role-specific instruction responses. The evaluations conducted in this thesis focus primarily on the role-specific component SPE.

Both LoRA-based variants improve over the base model on RoleBench SPE. The standard LoRA adapter improves the score from 20.23 to 24.78, an absolute increase of 4.55 points. The routed clustered-LoRA system further improves the score to 26.03, giving a 5.80 point improvement over the base model. This indicates that RoleBench-based adaptation helps the lightweight Phi-3 model produce responses that are closer

Table 5.2: Selected RoleBench paper references and Phi-3 results

System	Size / type	SPE $\uparrow$
LLaMA-2-7B-Chat reference	7B	20.2–21.4
Alpaca reference	7B	25.9–27.0
Vicuna reference	7B/13B	27.8–29.1
RoleGPT reference	GPT-based prompting	29.9–32.3
RoleLLaMA-7B reference	7B fine-tuned	25.7–38.1
RoleLLaMA2-13B-Chat reference	13B fine-tuned	48.8
Phi-3 base model	3.8B	20.23
Phi-3 + standard LoRA	3.8B + LoRA	24.78
Phi-3 + routed clustered LoRA	3.8B + 5 LoRA adapters	26.03

to the role-specific reference answers.

The clustered system also performs better than the single standard adapter. The improvement is smaller than the improvement from base to standard LoRA, but it suggests that grouping roles before adapter training can provide some additional benefit for in-domain role-specific generation.

5.1.1 Comparison with RoleBench Paper Results

The base Phi-3 model is close to the lower RoleBench paper references such as LLaMA-2-7B-Chat, despite being a smaller model. After LoRA adaptation, Phi-3 approaches the lower range of 7B role-playing references such as Alpaca and the role-generalization value of RoleLLaMA-7B. However, it remains clearly below stronger role-specialized systems such as RoleGPT and RoleLLaMA2-13B-Chat. This positioning is expected because the thesis uses a lightweight 3.8B backbone and parameter-efficient adaptation rather than full-scale role-playing model development.

Overall, RoleBench provides the clearest evidence that LoRA adaptation improves in-domain role-playing behaviour for the selected lightweight model.

5.2 CharacterBench Results

The CharacterBench is used as an external test bench in order to check whether adaptation based on RoleBench has been successful in a broad character customisation

Table 5.3: CharacterBench average scores for Phi-3-based models

Model / system	CharacterBench average $\uparrow$
Phi-3 base model	3.34
Phi-3 + standard LoRA	3.11
Phi-3 + routed clustered LoRA	3.28

context. This means CharacterBench does not train models but rather tests how well RoleBench-based adaptation works across different aspects of characters. Specifically, CharacterBench measures several aspects of character behavior: memory consistency, factual accuracy, boundary consistency (the amount of time taken before a character says something out of bounds), personality consistency, emotional response, morality, human-likeness and engagement [30]. Each score is normalised to a 5 point scale and then measured by Character Judge.

The base Phi-3 model achieves the highest CharacterBench average among the tested Phi-3 variants, with a score of 3.34. The routed clustered LoRA system reaches 3.28, which is close to the base model but still slightly lower. The standard LoRA adapter performs lowest, with an average score of 3.11.

The results show that the Rolebench was adapted (fine-tuned) for RoleBench. It seems this fine-tuning did not necessarily carry over into good performance at Character Bench. In other words, one possible reason for this poor performance may be that both RoleBench and Character Bench measure different kinds of “role” behavior. While RoleBench scores reward similarity to the specific answer given during training for each individual role; Character Bench uses a judge to rate many different kinds of character customization, from several different perspectives. Therefore, there is a possibility that adapting to be better at answering questions according to the RoleBench style will also cause loss of some generalizability or naturalness/fluency, or even some kind of judge preferred type of characterization when measuring at Character Bench.

5.2.1 Comparison with CharacterBench Paper Results

The base Phi-3 model achieves 3.34, which is above the CharacterGLM-6B reference mean of 3.20, but below than more stronger open source models such as Yi-1.5-

Table 5.4: Selected CharacterBench paper references and Phi-3 results

System	Size / type	Avg. zh	Avg. en	Mean
CharacterGLM-6B reference	6B	3.21	3.19	3.20
Yi-1.5-9B-Chat reference	9B	3.43	3.47	3.45
Mistral-7B-Chat reference	7B	3.50	3.55	3.53
GLM4-9B-Chat reference	9B	3.58	3.58	3.58
Llama3-8B-Instruct reference	8B	3.60	3.65	3.63
GPT-3.5-turbo reference	closed-source	3.66	3.72	3.69
GPT-4-1106 reference	closed-source	3.69	3.74	3.72
Claude-3-opus reference	closed-source	3.82	3.88	3.85
Phi-3 base model	3.8B	–	–	3.34
Phi-3 + standard LoRA	3.8B + LoRA	–	–	3.11
Phi-3 + routed clustered LoRA	3.8B + 5 LoRA adapters	–	–	3.28

9B-Chat, Mistral-7B-Chat, GLM4-9B-Chat, and Llama3-8B-Instruct. The routed clustered-LoRA system achieves 3.28, not going further much from the base model. The standard LoRA adapter achieves 3.11, which is lower than two of both the base model and the selected paper references. These results suggest that RoleBench adaptation did not improve CharacterBench performance, although the clustered adapter setup reduced the degradation relative to the single global adapter.

5.3 PersonaGym Results

PersonaGym evaluates persona-agent behaviour across five task categories: Expected Action, Toxicity Control, Linguistic Habits, Persona Consistency, and Action Justification [24]. Unlike RoleBench, PersonaGym does not use lexical overlap. Instead, it uses rubric-based evaluation and reports an overall PersonaScore.

The local PersonaGym benchmark used in this thesis contains 198 personas and 9,900 prompts. The model responses are generated first, converted into PersonaGym-compatible files, and then evaluated with the PersonaGym scoring pipeline.

The scoring in this thesis is accomplished by utilizing a version of `gpt-5-nano` as an evaluator (and thus different than the GPT based evaluator models described in the paper; specifically, as noted earlier GPT-4o). I selected to use `gpt-5-nano` due to the high computational cost associated with local PersonaGym evaluations – given that each response needs to be evaluated thousands of times. Thus, the scores can

Table 5.5: PersonaGym task-level results for Phi-3-based models

Task / metric	Base	Standard LoRA	Routed LoRA
Expected Action	3.74	3.34	3.58
Toxicity Control	4.92	4.84	4.89
Linguistic Habits	2.82	2.22	2.44
Persona Consistency	4.62	4.32	4.68
Action Justification	4.12	4.35	4.26
PersonaScore	4.04	3.81	3.97

not be compared directly to those obtained in the paper.

The base Phi-3 model achieves the highest overall PersonaScore among the tested variants, with 4.04. The routed clustered-LoRA system obtains 3.97, which is close to the base model, while the standard LoRA adapter obtains 3.81.

At the task level, the standard LoRA adapter improves Action Justification compared with the base model, but it decreases Expected Action, Linguistic Habits, Persona Consistency, and the overall PersonaScore. The routed clustered-LoRA system performs better than the standard adapter and even slightly exceeds the base model on Persona Consistency, with 4.68 compared with 4.62. However, it remains below the base model on the overall PersonaScore.

The results are consistent with our previous findings that LoRA adaptations are not uniformly beneficial across all of the persona-evaluation settings. On PersonaGym, there is a pre-existing advantage for the base model trained solely on the instructions, so we interpret the RoleBench adaptation as causing the model to become too narrowly focused on role-playing data instead of enhancing overall persona-agent behavior.

5.3.1 Comparison with PersonaGym Paper Results

The base Phi-3 model obtains a PersonaScore numerically close to the LLaMA-2-13B reference and above the Claude 3 Haiku reference. However, because the evaluator model differs from the original paper setup, this should be treated as an approximate comparison of score ranges rather than a direct claim of superiority.

The results have implications because they demonstrate that although the small

Table 5.6: Selected PersonaGym paper references and Phi-3 results

System	Size / type	PersonaScore $\uparrow$
Claude 3 Haiku reference	closed-source	3.64
LLaMA-2-13B reference	13B	3.98
GPT-3.5 reference	closed-source	4.38
LLaMA-2-70B reference	70B	4.39
LLaMA-3-8B reference	8B	4.49
Claude 3.5 Sonnet reference	closed-source	4.51
Phi-3 base model	3.8B	4.04
Phi-3 + standard LoRA	3.8B + LoRA	3.81
Phi-3 + routed clustered LoRA	3.8B + 5 LoRA adapters	3.97

(lightweight) Phi-3 base model may be less effective than the larger model at processing the PersonaGym task, it has a non-trivial PersonaScore. Nonetheless, adding the LoRA adapters did not enhance this overall score. Therefore, the evidence supports the interpretation that RoleBench adaptation enhances in-domain RoleBench SPE; however, it does not support the conclusion that RoleBench adaptation would generally lead to improved persona-agent performance, as measured by PersonaGym.

5.4 MPI/IPIP-Style Personality Evaluation

The MPI/IPIP-style evaluation measures personality-oriented response tendencies using Big Five-style questionnaire items. The evaluation follows the idea of the Machine Personality Inventory, where language models answer structured personality items and their responses are converted into OCEAN trait scores [13]. It is also related to IPIP-based work comparing model outputs with human personality baselines [15].

These results should not be interpreted as evidence that a model has a real personality. Instead, the scores are behavioural indicators showing how the model responds to personality-oriented prompts. Unlike RoleBench, CharacterBench, and PersonaGym, MPI/IPIP prompts do not offer any role or persona description to help route. The evaluation of clustered-LoRA in this case is done using a static clustered-adapter configuration and not dynamic routing per each questionnaire question.

LoRA adaptation changes the model’s OCEAN-style response profile. When comparing with the base model, the standard LoRA adapter increases Openness and

Table 5.7: MPI/IPIP-style OCEAN scores for Phi-3-based models

Trait	Base	Standard LoRA	Routed LoRA
Openness	3.72	4.08	4.02
Conscientiousness	3.58	3.48	3.56
Extraversion	4.13	3.56	3.47
Agreeableness	2.97	3.87	3.76
Neuroticism	3.12	2.58	2.82

Table 5.8: MPI reference results and Phi-3-based OCEAN scores

System	Size / type	O	C	E	A	N
Alpaca reference	7B	3.58	3.75	4.00	3.50	2.75
GPT-3.5 reference	closed-source	3.50	3.83	4.00	3.58	3.12
Human average	human baseline	3.44	3.60	3.41	3.66	2.80
Phi-3 base model	3.8B	3.72	3.58	4.13	2.97	3.12
Phi-3 + standard LoRA	3.8B + LoRA	4.08	3.48	3.56	3.87	2.58
Phi-3 + routed clustered LoRA	3.8B + 5 LoRA adapters	4.02	3.56	3.47	3.76	2.82

Agreeableness but at the same time reducing Extraversion and Neuroticism. The routed clustered-LoRA system shows a similar pattern: higher Openness and Agreeableness than the base model, lower Extraversion, and slightly lower Neuroticism.

The largest change is in Agreeableness. The base model scores 2.97, while the standard LoRA adapter scores 3.87 and the routed clustered system scores 3.76. This may indicate that RoleBench adaptation makes the model more cooperative or socially accommodating in questionnaire-style responses. At the same time, Extraversion decreases from 4.13 in the base model to 3.56 and 3.47 in the adapted variants. This suggests that role-playing adaptation does not simply make the model more expressive across all traits; instead, it shifts the trait profile in a more specific direction.

5.4.1 Comparison with MPI Reference Results

The base Phi-3 model resembles GPT-3.5 in Neuroticism and Extraversion but is lower in Agreeableness than both GPT-3.5 and the human average. After LoRA adaptation, the model’s Agreeableness moves closer to the human average, while its Extraversion becomes closer to the human average than the base model. The routed clustered-LoRA model is particularly close to the human average on Conscientiousness, Extraversion, Agreeableness, and Neuroticism, although its Openness remains higher.

Table 5.9: Summary of main results across evaluation settings

Metric	Base	Standard LoRA	Routed LoRA
RoleBench SPE $\uparrow$	20.23	24.78	26.03
CharacterBench avg. $\uparrow$	3.34	3.11	3.28
PersonaGym PersonaScore $\uparrow$	4.04	3.81	3.97
MPI Agreeableness	2.97	3.87	3.76
MPI Neuroticism	3.12	2.58	2.82

This doesn’t mean that LoRA makes the model more human like sound in a general psychological sense. However, it shows that RoleBench adaptation changes personality questionnaire response tendencies in a measurable way. For example, the adapted models produce a less extraverted and more agreeable profile than the base model.

5.5 Cross-Benchmark Comparison

The results across benchmarks show that the effect of LoRA adaptation depends strongly on the evaluation setting, Table 5.9.

The strongest positive result is observed on RoleBench. Both LoRA versions outperform the base model significantly, while the routed clustered version has been shown to perform the best. Therefore, based upon the results, we believe RoleBench adaptation was able to successfully adapt role specific answers to better fit the questions asked on the RoleBench dataset which was one of our training datasets.

On the other hand, we find that when evaluating performance on the external benchmarks (CharacterBench & PersonaGym), we have seen a more conservative view. When comparing all three models on the external benchmarks, we found that the base model tends to remain superior across most benchmarks compared to both adapted variants. Additionally, we see that the routed clustered LoRA variant tends to be much closer to the base model than the standard LoRA variant; therefore, we conclude that clustering may reduce some of the degradation associated with adapting globally, however, it still did not exceed the base model on either of the external benchmarks.

When evaluating the Personality Questionnaires using the MPI/IPIP style, we found evidence supporting the claim that adaptation alters questionnaire style behavioral tendencies (i.e., more agreeable and less neurotic). We also saw that Extraversion scores were moved toward the human average score listed in the MPI Reference Comparison. Thus, we feel confident in stating that RoleBench adaptation influences not only role-specific responses but also broader questionnaire-style behavioral tendencies.

5.6 Summary of Main Findings

The results support three main findings.

First, LoRA adaptation improves in-domain role-playing performance. RoleBench SPE increases from 20.23 for the base model to 24.78 for the standard LoRA adapter and 26.03 for the routed clustered-LoRA system. This confirms that parameter-efficient fine-tuning can improve a lightweight model on role-specific response generation.

Second, the improvement does not transfer uniformly to external persona benchmarks. On the full CharacterBench evaluation, the base model achieves 3.34, while the standard LoRA adapter and routed clustered-LoRA system achieve 3.11 and 3.28 respectively. On PersonaGym, the base model obtains 4.04, while the standard and routed LoRA variants obtain 3.81 and 3.97. These results indicate that RoleBench fine-tuning may specialize the model toward RoleBench-style role-playing without necessarily improving broader character or persona-agent behaviour.

Third, clustered LoRA is more effective than standard LoRA in this experimental setup. Across RoleBench, CharacterBench, and PersonaGym, the routed clustered-LoRA system outperforms the standard LoRA adapter. Although it does not always outperform the base model, it reduces the performance drop observed with the single global adapter and achieves the best RoleBench score. This suggests that clustered adapter specialization is a promising direction, but the current routing and training setup is not sufficient to guarantee improvements across all benchmarks.

5.7 Limitations of the Results

Several limitations should be considered when interpreting these results.

Beginning at the top of this section, the first point that will be made is that each of these benchmarks are evaluating unique aspects of personae behavior. Each of the four benchmarks evaluate personae behavior in somewhat unique ways: RoleBench specifically focuses on reward for role-specific lexical similarity; CharacterBench is focused on multi-dimensional character development and/or customization; PersonaGym is using rubric based evaluation to assess how well a given personae can exhibit agent-behavior, similar to what was shown in prior work; and MPI/IPIP style testing is simply a questionnaire-style assessment of trait tendency.

Therefore, if one benchmark shows some type of improvement, it does not automatically mean that another will also have improvement.

A second point that will be made is that the comparison to benchmark-papers' results is not a completely fair apples-to-apples comparison. In many cases, the authors of the papers being compared to are comparing two completely different models, model size(s), prompting strategy(ies), judge model(s), and even, in some cases, completely different benchmark implementations. This is particularly true when looking at results from PersonaGym, which the author of this thesis has had to implement using `gpt-5-nano` as a less expensive evaluator, whereas the authors of the original paper were able to use much stronger GPT-based judge models (e.g., GPT-4o).

A third point that will be made is that LoRA adaptation changes the response-distribution of a model, however it does so without adding an explicit memory mechanism. As such, any improvement in consistency or personae behavior seen due to LoRA adaptation should be viewed as the result of supervised adaptation rather than as evidence of long-term memory.

Lastly, clustered LoRA relies heavily upon the quality of both the clustering of role descriptions and the routing of adapters. That is, if a test-persona is poorly represented within any RoleBench role-cluster, then the routed-system will likely

route the input-test-persona to an adapter that has little relation to the intended-persona.

5.8 Summary

This chapter evaluated the base Phi-3 model, a typical RoleBench LoRA adapter, and a routed clustered-LoRA system on RoleBench, CharacterBench, PersonaGym, and MPI/IPIP-style testing. The findings indicate that the adaptation of LoRA enhances RoleBench SPE, and the routed clustered-LoRA system has the highest in-domain role-playing score. But the adapted models do not transfer better to external persona benchmarks than the base model, which suggests that the adaptation does not transfer to external persona benchmarks. This gives the final verdict as being mixed. The parameter-efficient fine-tuning of the lightweight model is useful in that it enables the lightweight model to learn RoleBench-style role-playing behaviour, but it does not guarantee more general improvements in character customization or persona-agent behaviour. Clustered LoRA has demonstrated advantages over the standard adapter, and it seems to be reliably beneficial in a wider range of persona evaluation environments.

Chapter 6

Discussion

The most significant implications of the research findings outlined within Chapter 5 are discussed in this chapter. The research question for this thesis was to evaluate if LoRa based adaptation would improve persona oriented behavior in a lightweight language model. In terms of the results, the findings indicated that there were some positive effects of adaptation on RoleBench performance when adapting to an in domain task; however, the results of the experiments also showed that adaptation was unable to produce consistent improvements in external benchmarks (such as CharacterBench and PersonaGym). Although the clustered LoRa approach produced superior performance compared with the baseline LoRa Adapter, the clustering technique was still not able to outperform the base model across all of the benchmarks.

6.1 Interpretation of the Main Results

The most significant improvement was found when using the RoleBench dataset. The base `microsoft/Phi-3-mini-4k-instruct` model generated an average SPE of 20.23 on the RoleBench dataset. This SPE value increased to 24.78 when using the standard LoRA adapter. Further still, use of the routed clustered-LoRA system resulted in an SPE value of 26.03. As RoleBench was the dataset used for training it demonstrates successful adaptation of the lightweight model to generate responses similar to those found in RoleBench, which are role-specific.

Improvement however, was not seen in the other datasets tested. For example, Microsoft’s base `Phi-3-mini-4k-instruct` model had an average score of 3.34, whereas both the standard LoRA adapter (3.11), and the routed-clustered-LoRA systems (3.28) were significantly lower than the base model when testing with the CharacterBench dataset. Similarly, when testing with PersonaGym, Microsoft’s base `Phi-3-mini-4k-instruct` model achieved an PersonaScore of 4.04, while the standard LoRA adapter (PersonaScore = 3.81), and the routed-clustered-LoRA systems (PersonaScore = 3.97) again fell short of the base model.

These test results demonstrate that while there was improvement in terms of adapting the model to generate role-specific responses within the domain of RoleBench; there was no general increase in persona-related behaviours beyond the specific domain tested by RoleBench.

There are many differences among the two benchmarking datasets. Benchmark tests provide evaluations of performance based upon what each individual data set measures. RoleBench evaluates how well a model can respond in ways that reflect the characteristics of a particular role, i.e., as measured by similarity of responses to roles. In contrast, CharacterBench uses metrics like fact accuracy, boundary consistency, emotional response, human-likeness, and engagement to evaluate models’ ability to create characters. Additionally, PersonaGym assesses agents’ ability to behave in ways consistent with personas through a variety of tasks including Expected Action, Linguistic Habits, Persona Consistency, Toxicity Control, and Action Justification.

6.2 Effect of Clustered LoRA

Clustered-LoRA system was more effective than the standard LoRA adapter in the key benchmarks. It also scored highest on RoleBench, 26.03, as opposed to 24.78 in the standard adapter. It also fared better on CharacterBench and PersonaGym against the base model. On CharacterBench, clustered LoRA had a score of 3.28 as compared to 3.11 of standard LoRA. It rated 3.97 on PersonaGym as opposed to 3.81. This implies that role clustering prior to the training of the adaptors minimized some

of the adverse effects of using the same global adaptor. RoleBench has heterogeneous roles, which have various styles and behavioural patterns. Adapters can be averaged together by a single adapter, or by clustering adapters can specialise on more coherent clusters of roles. Meanwhile, clustered LoRA did not always perform better than the base model on external benchmarks. It implies that clustering was helpful in comparison with basic LoRA, yet, it did not fully resolve the problem of transfer. It is a matter of the quality of clustering of roles-description and whether the test personality can be redirected to the corresponding cluster of roles.

6.3 Personality-Oriented Behaviour

The MPI/IPIP-style assessment revealed that the personality-oriented response tendencies of the model were altered due to the MPI/IPIP-style adaptation. The modified models were more amiable and less neurotic compared to the base model. An example is that Agreeableness went up in the base model from 2.97 and in the standard LoRA adapter to 3.87 and in the routed clustered-LoRA system to 3.76. The neuroticism dropped from 3.12 to 2.58 and 2.82 respectively. These modifications are not to be taken as indications that the model is has a real personality. Instead, they demonstrate how adaptation influences the manner in which the model is responsive to structured personality-style prompts. This proves to be helpful over the fact that it indicates that role-playing fine-tuning modifications do not merely benchmark-specific responses, but also wider response patterns.

6.4 Implications

The findings indicate that LoRA can be applied to specific persona adaption and particularly in cases where the task to be evaluated is nearest to the training data. The RoleBench enhancement demonstrates that a lightweight model can be trained to follow role-specific patterns of responding by parameter-efficient fine-tuning. Nonetheless, the findings further indicate adaptation beyond the training benchmark. Had

RoleBench only been considered the verdict would have been quite positive. CharacterBench and PersonaGym demonstrate that the modified models do not always generalise to more general persona and character assessment environments. This implies that persona adaptation may be benchmark specific. This comparison with paper results also indicates that lightweight models can be competitive with some lower reference systems, but are below stronger role-specialised or larger models. This is not surprising, as the thesis operates a 3.8B backbone as well as LoRA adaptation as opposed to more sizeable training or closed-source frontier models.

6.5 Limitations

There are a few limitations that should be taken into account. To start with, the benchmark comparisons are not the perfectly controlled since the original papers apply various models, prompts, judge models and evaluation implementations. This is more important in the case of PersonaGym because this thesis utilized `gpt-5-nano` as an evaluator with a low cost, whereas the original paper involved more powerful GPT-based judge models, such as GPT-4o. Second, the benchmarks are used to test various elements of persona behaviour. RoleBench employs reference-based Rouge-L, CharacterBench uses judge-based character dimensions, PersonaGym uses rubric-based persona-agent assessment, and MPI/IPIP-style testing measures questionnaire-style tendency to traits. Thus, a particular score cannot be regarded as a full-fledged measure of persona ability [3]. Third, the thesis considers adaptation by itself. It does not entail an external memory or retrieval mechanism. Consequently, improvement in consistency which is observed is due to fine tuning and not long term memory. Fourth, clustered LoRA relies on the quality of clustering and routing. When a test persona does not fit well into a RoleBench cluster, the routed system may select an adapter that only has a weak connection to the desired persona.

6.6 Future Work

The approach may be enhanced in various ways in future work. First, adapters could be trained on a more diverse mixture of persona datasets instead of RoleBench alone. It can enhance the transfer to such benchmarks as CharacterBench and PersonaGym. Second, the routing of clustered adapters might be enhanced. As an alternative to taking a single nearest cluster, soft routing, interpolating adapters, or routing under uncertainty can be used in future work. Third, the human judgment might be included. Scalable comparison can be effectively done through automatic benchmarks though persona quality is somewhat subjective and human judgments would offer another point of view. Fourth, it would be interesting to repeat the same experiment with other lightweight backbones to see whether the results obtained were specific to Phi-3 or general across compact LLMs. Fifth, the additional memory module or some memory augmented methods could be added for even better results in long term dialogues [7, 19].

6.7 Closing Perspective

Generally, the thesis would end in a moderate conclusion. Adaptation with LoRA enhances lightweight model on RoleBench, and clustered LoRA is better than single global adapter. But these profits do not always flow to CharacterBench or PersonaGym. Thus, parameter-efficient fine-tuning can be beneficial in targeted role-playing adaptation, but cannot be used alone to ensure that personas are consistent across different evaluation conditions.

Chapter 7

Conclusion

This paper examined how parameter efficient LoRa adaptation could affect persona orientation in a minimal language model. Microsoft’s `Phi-3 Mini 4K Instruct` was selected as the base model. Two other versions of this model were created through adaptation; one version was a standard LoRa adapter, which was trained on the RoleBench dataset. The second version was a routed cluster-LoRa system, which was also trained on clusters of role description datasets. These three versions were evaluated against each other using the RoleBench, CharacterBench, PersonaGym and an additional custom MPI/IPPI style personality evaluation based on previous psychological evaluations of LLMs [13, 15, 24, 27, 30].

This investigation’s primary research question was if LoRa-based adaptations would lead to better behavior in a lightweight model relative to their base version in terms of learning roles, playing roles, being consistent with characters, and having a personality based on personas. LoRa is an adaptable method that allows for parameter efficient adaptation by updating a low rank matrix of updates while maintaining the original model weights constant [11]. The results show that there are both positive and negative responses. There was a clear positive effect from LoRa adaptation when generating role-specific answers on the RoleBench. The SPE scores from the base model were at 20.23. SPE scores from the standard LoRa adapters were at 24.78. SPE scores from the routed cluster-LoRa systems were at 26.03. As such, there appears to be evidence that parameter efficient adaptation can allow a

compact model to learn to behave similarly to the types of behaviors shown in the training data related to role-playing.

However, the improvement from LoRA adaptation did not appear to be consistent across all benchmarks. When comparing the SPE scores of the Phi-3 mini 4K instruct model and its adaptations on CharacterBench (a benchmark that evaluates broad character customization and persona agent behavior), the base model scored higher than either of the two adapted versions, achieving average score of 3.34, compared to 3.11 for the standard LoRA adapter and 3.28 for the routed clustered-LoRA system. Similarly, when evaluating the models’ ability to perform well as agents based on their personalities and personas on PersonaGym, the base model again out performed its adapted versions, achieving a Persona Score of 4.04, compared to 3.81 for the standard LoRA adapter and 3.97 for the routed clustered-LoRA system. Because RoleBench is focused strictly on role-playing in domains similar to those seen in RoleBench, it may be reasonable to assume that improvements made possible through adaptation will result in better role-playing in those same domains. However, because the broader character customization and persona agent behavior represented in CharacterBench and PersonaGym respectively include many more aspects beyond just role-playing, these results indicate that improvements in domain specific role-playing do not necessarily translate into improving overall persona oriented behavior [24, 30].

A positive result was obtained from the use of the clustered-LoRA configuration compared with the standard global adapter. In comparison to the standard LoRA adapter, the routed clustered-LoRA systems were superior in terms of performance for RoleBench, CharacterBench and PersonaGym. It follows that there will be an improvement in the degree of over-specialisation or interference present in the heterogeneous role-playing data that is used during the training process of a single adapter due to the grouping of roles prior to this type of training. However, the clustered LoRA configurations did not produce consistent better performance than the baseline model for these evaluations. As such, clustering may be useful; however, it cannot alone ensure a wide range of generalisations.

Using the MPI/IPIP-type method, which assesses whether adaptation affects

personality-oriented response characteristics, demonstrated that LoRA-adaptation resulted in significant changes in response style based on personality. Specifically, the adapted models were found to be more agreeable, less neurotic, and had a higher extraverted score (i.e., closer to the mean of humans) than the base model. The results obtained using this methodology do not demonstrate that the model possesses a ‘real’ personality. Rather, the purpose of this type of evaluation (similar to that of MPI and IPIP-type methods) is to provide evidence that fine-tuning on specific role-playing datasets causes a shift in how a model responds to prompts presented in a structure designed to elicit personality-style responses [13, 15].

In summary, the findings support a moderate conclusion. Adaptation through LoRA is beneficial for improving lightweight models on targeted role-playing tasks, particularly when evaluating performance on tasks that are close to those used for training. However, adapting to a single role-playing dataset does not imply an improved generalisation capability for all benchmarking tests focused on persona-oriented tasks.

A number of constraints can be mentioned. First, the benchmark comparisons to original papers are approximate since the papers employ different models, prompts, judge models and evaluation settings. Second, other assessments use automatic judges, which can cause scoring bias or noise. Third, the thesis only involves adaptation but not retrieval-augmented memory and long-term dialogue memory. Hence, the outcomes are to be interpreted as the impact of LoRA-based fine-tuning as opposed to the solution to the long-term memory of personas. Lastly, the clustered adapter system relies on the effectiveness of role clustering and routing, which are not necessarily effective at selecting the most appropriate adapter to a new persona.

There are multiple ways that future research can build upon the findings presented in this paper. For instance, the adapters could be trained using a wider range of persona-centric corpora than just RoleBench to boost performance on CharacterBench and PersonaGym. Additionally, clustered-adapter routing can be improved, for instance via soft routing, adapter interpolation, or choosing based on uncertainty. It would also be beneficial to conduct human evaluations because persona quality is

somewhat subjective, and some aspects of it might not be measurable automatically. Lastly, the proposed method can be evaluated using different lightweight architectures to verify that the results achieved are architecture-specific or universal to compact LLMs.

Overall, this thesis demonstrates how parameter-efficient fine-tuning could enhance a lightweight language model in role-playing assessment, with clustered LoRA outperforming its non-clustered counterpart in such a scenario. On the other hand, the findings highlight the limitations of the technique by illustrating that an improvement in performance in the RoleBench benchmark does not guarantee similar success in other persona and character benchmarks. In summary, LoRA adaptation is a viable method for lightweight persona modeling but should be subjected to thorough testing on multiple benchmarks for general persona-based dialogue.

Bibliography

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.
- [2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [3] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024.
- [4] Tapan Kumar Das and Ayan Biswas. Fine-tuning and optimizing lightweight llms for domain-specific use cases. In *2025 Innovations in Power and Advanced Computing Technologies (i-PACT)*, pages 1–5, 2025.
- [5] Jaap J. A. Denissen, Rinie Geenen, Christopher J. Soto, Oliver P. John, and Marcel A. G. van Aken. The big five inventory–2: Replication of psychometric properties in a dutch adaptation and first evidence for the discriminant predictive validity of the facet scales. *Journal of Personality Assessment*, 102(3):309–324, 2020. PMID: 30638406.
- [6] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jin Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023.
- [8] Lewis Goldberg. An alternative “description of personality”: The big-five factor structure. *Journal of Personality and Social Psychology*, 59:1216–1229, 12 1990.
- [9] Lewis R. Goldberg, John A. Johnson, Herbert W. Eber, Robert Hogan, Michael C. Ashton, C. Robert Cloninger, and Harrison G. Gough. The international personality item pool and the future of public-domain personality

- measures. *Journal of Research in Personality*, 40(1):84–96, 2006. Proceedings of the 2005 Meeting of the Association of Research in Personality.
- [10] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA, 2022. Curran Associates Inc.
 - [11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
 - [12] Linmei Hu, Xinyu Zhang, Dandan Song, Changzhi Zhou, Hongyu He, and Liqiang Nie. Efficient and effective role player: A compact knowledge-grounded persona-based dialogue model enhanced by llm distillation. *ACM Trans. Inf. Syst.*, 43(3), February 2025.
 - [13] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. Evaluating and inducing personality in pre-trained language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS ’23, Red Hook, NY, USA, 2023. Curran Associates Inc.
 - [14] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, et al. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
 - [15] Lawrence J. Klinkert, Steph Buongiorno, and Corey Clark. Evaluating the efficacy of llms to emulate realistic human personalities. In *Proceedings of the Twentieth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE ’24. AAAI Press, 2024.
 - [16] Pranjal Kumar. Large language models (llms): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57, 2024.
 - [17] Ang Li, Haozhe Chen, Hongseok Namkoong, and Tianyi Peng. Llm generated persona is a promise with a catch. *arXiv preprint arXiv:2503.16527*, 2025.
 - [18] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. A persona-based neural conversation model. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany, August 2016. Association for Computational Linguistics.
 - [19] Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. Augmented language models: a survey. 2023.

- [20] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, et al. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.
- [21] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. *ArXiv*, abs/2005.00247, 2020.
- [22] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. AdapterHub: A framework for adapting transformers. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online, October 2020. Association for Computational Linguistics.
- [23] Seyed Mahmoud Sajjadi Mohammadabadi, Burak Cem Kara, Can Eyupoglu, Can Uzay, Mehmet Serkan Tosun, and Oktay Karakuş. A survey of large language models: Evolution, architectures, adaptation, benchmarking, applications, challenges, and societal implications. *Electronics*, 14(18), 2025.
- [24] Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. PersonaGym: Evaluating persona agents and LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China, November 2025. Association for Computational Linguistics.
- [25] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [27] Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

- [28] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*, 2024.
- [29] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [30] Jinfeng Zhou, Yongkang Huang, Bosi Wen, Guanqun Bi, Yuxuan Chen, Pei Ke, Zhuang Chen, Xiyao Xiao, Libiao Peng, Kuntian Tang, Rongsheng Zhang, Le Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. Characterbench: Benchmarking character customization of large language models. 2024.