A Controlled Study of Parameter-Efficient Fine-Tuning in Cold-Start Recommendation

A Major Project Report
Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

DATA SCIENCE

by

RAJAT YADAV
24/DSC/18

Under the Supervision of

Dr. Rahul
Assistant Professor, Department of Software Engineering
Delhi Technological University

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2026

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my supervisor, Dr. Rahul for the continuous guidance, encouragement, and valuable suggestions throughout the course of this thesis work. Her/His insights have been instrumental in shaping the direction and quality of this research.

I also thank the faculty members of the Department of Software Engineering for their academic support and encouragement. Special thanks to the members of the Data Science program for creating a collaborative research environment.

I am grateful to my family and friends for their unwavering support and motivation throughout this challenging project. Their patience and understanding have been invaluable during the demanding periods of implementation, experimentation, and thesis writing.

Place: Delhi

Date: 06/07/26

Rajat

RAJAT YADAV.

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I hereby declare that the report work entitled

A controlled study of Parameter-Efficient Fine-Tuning in cold start Recommendation

submitted in partial fulfillment of the requirements for the award of the degree of Master of Technology in Data Science, in the Department of Software Engineering, Delhi Technological University, is a bona fide record of the research work carried out by me during the period from Aug 2024 to May 2026 under the supervision of Dr. Rahul.

I further declare that the work presented in this thesis has not been submitted, either in part or in full, for the award of any other degree, diploma, or fellowship of this or any other institution or university.

Rajat
Candidate's Signature

This is to certify that the candidate has incorporated all the corrections and suggestions recommended by the examiners, and the declaration made by the candidate is true and correct to the best of our knowledge and belief.

Signature of Supervisor(s)

Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CERTIFICATE

This is to certify that the Project Dissertation titled "A controlled Study of Parameter-efficient Fine-Tuning in cold start Recommendation", submitted by Rajat Yadav, Roll No. 24/DSC/10, Department of Software Engineering, Delhi Technological University, Delhi, in partial fulfillment of the requirement for the award of the degree of Master of Technology (Data Science), is a record of the project work carried out by the student under my supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 06/07/26

Dr. Rahul
Assistant Professor.
Department of Software Engineering

ABSTRACT

Cold-start recommendation remains one of the most persistent and practically consequential challenges in the design and deployment of personalised information retrieval systems at scale. When users arrive on a platform with no prior interaction history, or when new items are introduced to a catalogue before any user has engaged with them, conventional collaborative filtering and matrix factorisation methods suffer severe and often catastrophic degradation in their ranking quality, because these methods derive their predictive power entirely from the density of the observed user–item interaction matrix. The emergence of Large Language Model (LLM) based recommender systems has opened a qualitatively different path toward solving this problem: by encoding user preference signals and item characteristics as structured natural language prompts, and by leveraging the rich semantic world knowledge encoded in billions of pretrained model parameters, LLM-based recommenders can reason about relevance and preference even in the near-complete absence of interaction data. This capability has made LLM-based recommendation a topic of intense research interest, and has driven substantial investment in their deployment across commercial recommendation platforms.

However, the practical utility of LLM-based recommenders in cold-start settings is fundamentally contingent upon an efficient and effective adaptation strategy. A general-domain LLM, even a very capable one, must be aligned to the specific vocabulary of preferences, the item ontology, and the interaction patterns of the target recommendation domain. Full fine-tuning of billion-parameter models for this purpose is computationally prohibitive, memory-intensive, and poorly suited to the dynamic, rapidly evolving user–item distributions that characterise real recommendation platforms. Parameter-Efficient Fine-Tuning (PEFT) methods have been developed to address this adaptation bottleneck, selectively updating a small fraction of model parameters while keeping the vast majority of the pretrained weights frozen. Despite their growing importance, however, no prior work has provided a rigorous, unified, and controlled comparison of PEFT methods specifically in the cold-start recommendation setting.

This thesis addresses that gap directly. We design and execute a comprehensive benchmark framework for the fair comparison of three major PEFT families — Low-Rank Adaptation (LoRA), prompt-based tuning (soft prompt tuning and prefix tuning), and adapter-based tuning, including the specialised variant of Late Adapter Tuning (LAT) — under both the user cold-start and item cold-start regimes. The framework holds the backbone model, training optimiser, evaluation protocol, and dataset preprocessing pipeline constant across all compared methods, ensuring that observed performance differences are attributable to the adaptation mechanism alone. Six benchmark datasets are used for evaluation: the MIND news recommendation dataset, the Amazon product domains of Games, Toys, and Office, and the Netflix streaming dataset, evaluated under the all-ranking protocol using Recall@K, NDCG@K, and Hit Ratio@K metrics.

The central empirical finding of this thesis is that adaptation performance in cold-start recommendation is governed not primarily by the number of parameters updated, but by where in the model those updates are applied. Late Adapter Tuning, which confines adapter modules exclusively to the upper transformer layers of the pretrained backbone, achieves performance at or above full fine-tuning on multiple benchmarks while updating only 2.37% of total model parameters, reducing GPU memory consumption by 76.5% relative to full fine-tuning, and attaining the highest inference throughput among all compared methods. Prompt-based tuning, while offering the smallest possible parameter overhead at less than 0.01% of total parameters, proves to be expressively limited in highly personalised and long-tail recommendation settings where surface-level input conditioning is insufficient to redirect the model's internal representations toward task-specific preference patterns. LoRA occupies a well-balanced intermediate position, delivering near-full-fine-tuning accuracy with 0.22% of total parameters and strong optimisation stability. A structured cold-start robustness analysis reveals an important asymmetry between user- and item-cold-start failure modes: item cold-start is primarily a knowledge availability problem best addressed by retrieval-augmented methods, while user cold-start performance scales monotonically with interaction count and is governed by the model family and adaptation quality. Together, these findings establish a unifying principle that reframes cold-start recommendation from a data compression problem into a representation alignment problem, in which the goal of PEFT is to identify the minimal and most task-relevant parameters in the upper model layers and update them with maximal precision.

CONTENTS

Abstract **i**

List of Tables **iii**

List of Figures **iv**

List of Abbreviations **v**

1 Introduction **1**

1.1 The Cold-Start Problem: Origins and Significance 2

1.2 Large Language Models as a New Paradigm for Cold-Start Recommendation 3

1.3 Parameter-Efficient Fine-Tuning: Bridging Capability and Practicality 4

1.4 Research Questions and Objectives 5

1.5 Principal Contributions 5

1.6 Organisation of the Thesis 6

2 Technical Background **7**

2.1 Transformer Architecture and Large Language Models 7

2.2 Low-Rank Adaptation 8

2.3 Prompt-Based Tuning Methods 9

2.4 Adapter-Based Tuning Methods 10

2.5 The Cold-Start Recommendation Task: Formal Definition 11

2.6 Evaluation Metrics 11

3 Related Work and Literature Review **13**

3.1 LLM-Based Recommendation Systems 13

3.2 Parameter-Efficient Fine-Tuning Methods 15

3.3 Cold-Start Recommendation Approaches 16

3.4 Gap Analysis and Positioning of This Work 18

4 Methodology **21**

4.1 Problem Formulation 21
4.2 Parameter-Efficient Adaptation Configurations 22
4.2.1 Low-Rank Adaptation Configuration 23
4.2.2 Prompt-Based Adaptation Configurations 23
4.2.3 Adapter-Based Adaptation Configurations 24
4.3 Datasets 25
4.4 Experimental Setup 26
4.5 Baseline Systems 27
4.6 Design Rationale 27

5 Experimental Results and Discussion 29
5.1 Main Performance Results 29
5.2 Parameter Efficiency versus Adaptation Effectiveness 31
5.3 Cold-Start Robustness Analysis 33
5.4 Ablation Study 37
5.5 Efficiency and Practical Deployment Considerations 38
5.6 Synthesis and Theoretical Interpretation 39

6 Conclusion and Future Work 41
6.1 Conclusion 41
6.2 Future Work 42

LIST OF TABLES

3.1 Systematic Comparison of Related Works on Key Evaluation Dimensions. ✓= fully addressed, × = not addressed, P = partial coverage only. 19

5.1 Main Performance Comparison: RecLM Backbone Ablation and ColdRAG Cold-Start Results across Six Datasets. R@K = Recall@K, N@K = NDCG@K. Values for Games, Toys, and Office domains are reported as percentages (×100) following the convention in the Amazon product recommendation literature. 30

5.2 Parameter Efficiency vs. Accuracy for PEFT Methods. Backbone: RoBERTa-large (355M parameters). Bold values indicate the best performance per metric column. Speed is measured in tokens per millisecond (tok/ms). Memory consumption is reported in gigabytes (GB) during training. 32

LIST OF FIGURES

4.1 Schematic of the LLM-based cold-start recommendation pipeline. A new user with at most three prior interactions (user cold-start) and a candidate item described by its textual metadata (item cold-start) are processed through a structured natural language prompt. The prompt is passed to a frozen LLM backbone augmented with a PEFT adaptation module (LoRA, prompt tuning, or adapter). The model produces a relevance score for ranking, and the ranked list is evaluated using Recall@K, NDCG@K, and HR@K under the all-ranking protocol. . . 22

5.1 Parameter efficiency vs. accuracy trade-off across all PEFT methods evaluated on RoBERTa-large (355M parameters). The horizontal axis shows the percentage of trainable parameters relative to the full model on a logarithmic scale. The vertical axis shows the average accuracy across MPQA, MR, SUBJ, and TREC benchmarks. The LAT (Full) configuration occupies the dominant position in the trade-off landscape, achieving the highest average accuracy at 2.37% of total parameters. Full fine-tuning (100%) achieves lower average accuracy than LAT despite using 42× more trainable parameters. Prompt Tuning has the lowest parameter count but also the lowest accuracy. LoRA and the standard Adapter occupy intermediate positions, with LoRA slightly outperforming the adapter. LAT-down, despite identical parameter count to LAT (Full), achieves substantially lower accuracy, confirming that layer placement rather than parameter count is the determinant of adaptation quality. 34

5.2 Recommendation performance (Recall@10) as a function of user interaction history length (number of prior interactions available at inference time) for five representative recommendation models: RecLM, ColdRAG, LLMTreeRec, LightGCN, and SGL. All models show monotonically increasing performance as more interactions become available. LLM-based methods (RecLM, ColdRAG, LLMTreeRec) maintain substantially higher performance at very low interaction counts (≤ 3) than collaborative filtering baselines (LightGCN, SGL), with LightGCN and SGL showing near-zero performance at 0–1 interactions. The performance gap between LLM-based and collaborative filtering methods narrows as interaction count increases beyond 5–6 interactions, where collaborative filtering begins to accumulate sufficient signal to exploit its graph-based propagation mechanism effectively. . 36

LIST OF ABBREVIATIONS

API	Application Programming Interface
CF	Collaborative Filtering
DCG	Discounted Cumulative Gain
GPU	Graphics Processing Unit
GCN	Graph Convolutional Network
HR	Hit Ratio
LAT	Late Adapter Tuning
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MF	Matrix Factorisation
MIND	Microsoft News Dataset
NDCG	Normalised Discounted Cumulative Gain
NLP	Natural Language Processing
PEFT	Parameter-Efficient Fine-Tuning
RAG	Retrieval-Augmented Generation
ReLU	Rectified Linear Unit
SGL	Self-supervised Graph Learning
SVD	Singular Value Decomposition
tok/ms	Tokens per Millisecond
TREC	Text REtrieval Conference

Keywords: Cold-start recommendation, Large language model recommenders, Late adapter tuning, Low-rank adaptation, Parameter-efficient fine-tuning, Prompt-based learning, Retrieval-augmented recommendation, Sparse interaction modelling.

CHAPTER 1

INTRODUCTION

Recommender systems occupy a position of extraordinary practical importance in the contemporary digital economy. Every major digital platform — whether an e-commerce marketplace, a streaming service, a news aggregator, a social network, or a professional productivity tool — relies upon the capacity to rank and surface relevant items from catalogues of potentially millions of candidates in response to individual user preferences. The quality of this ranking determines not only user satisfaction and retention but, in commercial settings, also revenue, discovery of new content and products, and the fairness of exposure to producers and publishers. The challenge of building recommender systems that perform well across the enormous diversity of users, items, and interaction contexts they must serve is one of the central problems of applied machine learning.

The foundational approach to this challenge, collaborative filtering, rests on the observation that users who have agreed in the past tend to agree in the future. By representing users and items as points in a shared latent space learned from the pattern of observed interactions, collaborative filtering methods can predict which items a user is likely to prefer based on the preferences of similar users. Matrix factorisation, the most mathematically elegant realisation of this idea, decomposes the observed interaction matrix into the product of low-rank user and item embedding matrices, and the learned embeddings are used to score unobserved user-item pairs. This family of methods dominated the recommender systems research landscape for over a decade and continues to serve as the primary comparison baseline against which newer approaches are evaluated [1].

Yet the very mechanism that gives collaborative filtering its power is also the source of its most consequential limitation. Collaborative filtering is inductive in nature: it learns entirely from the pattern of observed interactions, and its predictive power degrades directly with the sparsity of those observations. Users who are new to a platform, and items that have only recently been added to a catalogue, present the system with an irreducible problem. In the absence of interaction data, the collaborative filter cannot place the user or item in the learned latent space, and any recommendation it

produces is essentially uninformed. This failure mode, known as the cold-start problem, is not an edge case or a minor inconvenience: it is a systematic and pervasive challenge that affects every new user at the moment of their highest vulnerability to churn, and every new item at the moment of its greatest need for discoverability.

1.1 The Cold-Start Problem: Origins and Significance

The cold-start problem was first identified as a formal research challenge in the early literature on collaborative filtering and has attracted sustained attention ever since. Its significance is not merely theoretical but deeply practical. In a streaming service, a new release must be surfaced to users who are likely to enjoy it before it accumulates enough ratings to be recommended by collaborative filtering; the window of opportunity for initial discovery is narrow and commercially critical. In an e-commerce platform, a new seller listing a product for the first time depends on the system's capacity to match that product to appropriate buyers without the benefit of historical click data. In a news recommendation context, where articles are produced at high velocity and their relevance window is measured in hours, the cold-start problem is essentially permanent: virtually every article the system must recommend is, at the time of recommendation, new.

Classical approaches to mitigating the cold-start problem have followed two broad strategies. The first is to enrich the interaction signal through auxiliary data: demographic information about users, attribute metadata about items, contextual signals such as the time of day or the device being used, and graph-structured data that captures social or co-purchase relationships. Content-based filtering, which recommends items similar to those a user has previously engaged with based on item feature similarity rather than collaborative signal, is the most widely deployed instantiation of this approach. The second strategy is to learn a transferable initialisation through meta-learning: rather than learning a single fixed model for all users, meta-learning methods learn an optimisation procedure that can adapt rapidly to a new user or item from very few examples, treating the cold-start problem as a few-shot learning task [2]. Both strategies have proven effective in controlled experimental settings, but both also have significant limitations: content-based filtering requires rich, structured, and discriminative item features; meta-learning requires careful task design and can be brittle under distribution shift.

1.2 Large Language Models as a New Paradigm for Cold-Start Recommendation

The emergence of Large Language Models (LLMs) trained on web-scale corpora of human-generated text has introduced a qualitatively different approach to the cold-start challenge. These models, having been exposed during pretraining to an extraordinarily broad range of factual, contextual, relational, and aesthetic information about the world, encode within their parameters a rich semantic understanding of entities, attributes, preferences, and relationships that is not available to any interaction-based model. When a user's preference history is expressed in natural language and concatenated with an item description as a structured prompt, an LLM can reason about the relevance of the item to the user's expressed interests without requiring any collaborative signal from an interaction matrix. It draws on its parametric knowledge of the domain, genre, style, and typical patterns of human preference, and produces a relevance assessment that is grounded in semantic understanding rather than statistical co-occurrence.

This paradigm shift, which reframes recommendation as a conditional text generation or instruction-following task rather than a matrix completion task, has demonstrated remarkable capability in low-resource and zero-shot settings. Approaches such as LLM-Rec [3] and the broader class of instruction-tuned recommendation frameworks have shown that pretrained language models can substantially outperform collaborative filtering baselines when interaction data is severely limited. The critical insight is that the LLM brings a rich prior over the space of human preferences that is derived not from any single platform's interaction data but from the full breadth of human-generated text about the world, and this prior proves to be directly useful for recommendation in cold-start settings even without task-specific training.

However, the practical deployment of LLM-based recommenders in cold-start settings is not without significant challenges of its own. The most fundamental of these is the cost of adaptation. A general-purpose LLM pretrained on a broad web corpus encodes world knowledge in a form that is not directly optimised for any specific recommendation domain. The particular vocabulary of user preferences, the idiosyncratic ontology of item categories, and the characteristic patterns of user-item affinity in a given platform are not necessarily well-represented in the pretrained parameter space. Achieving strong recommendation performance typically requires some form of fine-tuning that brings the model's representations into closer alignment with the task-specific preference distribution [4]. But full fine-tuning of a model with hundreds of millions or billions of parameters is computationally prohibitive in most practical

recommendation settings, where models must be updated frequently as the user-item distribution evolves and computational budgets are constrained.

1.3 Parameter-Efficient Fine-Tuning: Bridging Capability and Practicality

Parameter-Efficient Fine-Tuning (PEFT) methods represent the research community's answer to the adaptation cost problem. Rather than updating every parameter of the pretrained model during task-specific training, PEFT methods selectively modify a small and carefully chosen subset of parameters, either by injecting lightweight additional modules into the model architecture or by constructing compact parameterisations of the weight updates, while keeping the vast majority of the pretrained weights frozen. The result is a training procedure that can be several orders of magnitude less expensive than full fine-tuning in terms of both computational cost and memory consumption, while preserving — and in some configurations improving — the representational quality of the pretrained backbone for the downstream task.

Three families of PEFT methods have emerged as the most widely studied and practically deployed in the natural language processing research community. Low-Rank Adaptation, or LoRA [5], represents weight updates as the product of two low-rank matrices, exploiting the theoretical observation that the weight changes induced by fine-tuning have low intrinsic dimensionality. Prompt-based tuning methods, including soft prompt tuning and prefix tuning [6], introduce small sets of learnable tokens into the input sequence or the attention mechanism, conditioning the model's computation on task-specific signals without touching any of its pretrained parameters. Adapter-based methods [7] insert small, expressive bottleneck modules directly into the transformer blocks of the pretrained model, enabling targeted representational updates at specific architectural positions. Each of these three families reflects a different hypothesis about where in the model adaptation is needed and how it should be parameterised, and each exhibits a characteristic profile of trade-offs among parameter count, representational capacity, optimisation stability, and inference efficiency.

Despite the rapid growth of the PEFT literature and the widespread deployment of these methods across natural language processing tasks, their comparative behaviour in the specific and demanding context of cold-start recommendation has not been rigorously examined. Studies evaluating PEFT for recommendation tasks have typically evaluated a single method in isolation, varied the backbone model alongside the adaptation strategy, used inconsistent evaluation protocols, or focused on warm-start settings where users have substantial interaction histories. The result is a literature that

is fragmented, difficult to interpret, and of limited practical guidance for a practitioner seeking to choose an adaptation strategy for a cold-start recommendation deployment. This thesis directly addresses this gap.

1.4 Research Questions and Objectives

The research conducted in this thesis is motivated by a single central question: what is the relative effectiveness of different PEFT approaches, in terms of ranking performance, parameter efficiency, memory efficiency, and robustness to interaction sparsity, when applied to LLM-based recommenders in user- and item-cold-start settings? This central question gives rise to five specific research objectives that together define the scope and structure of the empirical investigation.

The first objective is to design and execute a rigorous, controlled experimental framework for the comparison of LoRA, prompt-based tuning, and adapter-based tuning under both the user cold-start and item cold-start regimes, holding the backbone architecture, training optimiser, context length, and evaluation protocol constant across all methods. This controlled design is the methodological foundation upon which all subsequent conclusions rest, and its careful execution is itself a contribution of the thesis. The second objective is to characterise the trade-offs among adaptation strategies across the four key dimensions of ranking accuracy on standard metrics, trainable parameter count as a percentage of total model parameters, peak GPU memory consumption during training, and inference throughput in tokens per millisecond. The third objective is to investigate the cold-start robustness of each adaptation approach as a function of interaction sparsity, examining how performance changes as the available interaction signal is reduced from a moderate few-shot level toward zero. The fourth objective is to conduct a structured ablation study that isolates the contributions of individual architectural choices within the best-performing configuration, establishing the structural requirements of effective adaptation. The fifth objective is to synthesise the findings into actionable design principles for practitioners and to identify the most productive directions for future research.

1.5 Principal Contributions

The principal contributions of this thesis are four in number. First, a unified and controlled benchmark framework is introduced for the evaluation of PEFT methods in cold-start recommendation, addressing a gap in the existing literature that has been a persistent barrier to reliable comparative analysis. This framework is notable for its

strict experimental controls, its use of the all-ranking evaluation protocol across all experiments, and its coverage of both user and item cold-start regimes across six datasets of diverse domain characteristics. Second, a comprehensive empirical evaluation is reported that establishes Late Adapter Tuning as a high-performance adaptation strategy whose superiority over full fine-tuning on multiple benchmarks — achieved with only 2.37% of total model parameters — establishes the principle of adaptation locality as a fundamental determinant of cold-start recommendation performance. Third, a structured analysis of cold-start robustness reveals the qualitative asymmetry between user- and item-cold-start failure modes and identifies retrieval-augmented adaptation as the most effective strategy for item cold-start in sparse product domains. Fourth, ablation experiments establish the architectural requirements of the LAT configuration and confirm that the advantage of upper-layer adaptation is not a consequence of adapter expressiveness per se but of the locality of the adaptation relative to the hierarchical structure of transformer representations.

1.6 Organisation of the Thesis

Chapter 2 provides the technical background necessary to understand the design decisions and evaluation choices in the thesis, covering the transformer architecture, the three principal PEFT families, the formal definition of the cold-start recommendation task, and the evaluation metrics employed throughout. Chapter 3 presents a comprehensive literature review organised around three thematic streams: LLM-based recommendation systems, parameter-efficient fine-tuning methods in natural language processing, and cold-start recommendation approaches, concluding with a systematic gap analysis. Chapter 4 describes the experimental methodology in full detail, including the problem formulation, the adaptation configurations, the datasets, the evaluation setup, and the design rationale. Chapter 5 presents and interprets the experimental results across four sub-studies: the main performance comparison, the parameter efficiency analysis, the cold-start robustness analysis, and the ablation investigation. Chapter 6 concludes the thesis with a synthesis of the principal findings and a detailed discussion of future research directions.

CHAPTER 2

TECHNICAL BACKGROUND

This chapter establishes the technical foundations necessary to understand the experimental design, the adaptation strategies under comparison, and the evaluation methodology employed throughout this thesis. The presentation begins with the transformer architecture that underlies all modern large language models, proceeds through the three families of parameter-efficient fine-tuning that form the central subject of the comparative investigation, and concludes with the formal definition of the cold-start recommendation task and the evaluation metrics used to quantify performance.

2.1 Transformer Architecture and Large Language Models

The transformer architecture, which today underpins virtually every state-of-the-art large language model, is built around the multi-head self-attention mechanism. Given an input sequence of n token embeddings $X \in \mathbb{R}^{n \times d}$, each transformer block computes, for each attention head h, a set of query, key, and value projections $Q = XW_Q^{(h)}$, $K = XW_K^{(h)}$, $V = XW_V^{(h)}$, where $W_Q^{(h)}, W_K^{(h)}, W_V^{(h)} \in \mathbb{R}^{d \times d_h}$ are learned projection matrices. The attention output for head h is then:

$$\mathrm{Attn}^{(h)}(X) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right) V \quad (2.1)$$

The outputs of all heads are concatenated and projected back to the model dimension d through a learned output projection $W_O \in \mathbb{R}^{Hd_h \times d}$, where H is the number of attention heads. This global attention mechanism enables every token to attend to every other token in the sequence regardless of their relative positions, giving the transformer an inductive capacity for modelling long-range dependencies that was substantially superior to the sequential recurrent architectures it supplanted. Each transformer block completes the attention sublayer with a position-wise feed-forward sublayer consisting of two linear transformations with a nonlinear activation between them, with layer normalisation and residual connections applied around both sublayers. The depth of the transformer stack — the number of such blocks stacked sequentially — is the primary

determinant of the model's capacity for hierarchical abstraction.

Large language models trained through next-token prediction on web-scale corpora of hundreds of billions of tokens acquire, as an emergent consequence of this training objective, a broad and integrated understanding of factual knowledge, linguistic pattern, commonsense reasoning, and domain expertise accumulated from the full breadth of human-generated text. This knowledge is not stored in any explicit symbolic structure but is implicitly encoded in the billions of learned parameters of the attention projection matrices and feed-forward layers, distributed across the entire model in a form that is not directly interpretable but is powerfully effective for a wide range of downstream tasks. The critical limitation of this parametric architecture, and the one that motivates the present thesis, is that this knowledge is frozen at the conclusion of pretraining and cannot be updated without retraining, while the model's internal representations may not be well-aligned with the specific preference vocabulary and interaction patterns of any particular recommendation domain.

Research into the hierarchical organisation of transformer representations has consistently found that lower layers of the transformer stack encode broadly transferable syntactic and surface-level linguistic features, while upper layers encode progressively more abstract, task-specific, and semantic representations [8]. This hierarchical structure has direct implications for the design of parameter-efficient fine-tuning strategies: adaptations applied to the upper layers are more likely to modify the task-relevant semantic representations without disrupting the broadly useful lower-layer features, while adaptations applied uniformly across all layers risk damaging the broadly transferable representations that constitute the foundation of the model's general-purpose competence.

2.2 Low-Rank Adaptation

Low-Rank Adaptation (LoRA) is grounded in a well-supported theoretical observation about the nature of fine-tuning: the changes in weight matrices induced by fine-tuning a large pretrained model on a specific downstream task tend to have a much lower intrinsic dimensionality than the full parameter dimensionality of those matrices. This observation suggests that the information content of fine-tuning can be captured by a much smaller set of parameters than full fine-tuning uses, without significant loss of performance. LoRA operationalises this insight by parameterising the weight update ΔW for any target weight matrix $W \in \mathbb{R}^{d \times k}$ as the product of two low-rank matrices:

$$\Delta W = BA, \quad B \in \mathbb{R}^{d \times r}, \quad A \in \mathbb{R}^{r \times k}, \quad r \ll \min(d, k) \tag{2.2}$$

During fine-tuning, only the matrices A and B are updated by gradient descent; the original weight matrix W is kept frozen throughout. The matrix A is initialised with random Gaussian entries scaled to a small standard deviation, while B is initialised to all zeros, which ensures that at the very start of fine-tuning the effective weight update $\Delta W = BA = 0$, preserving the pretrained model's behaviour at initialisation. At inference time, the adapted weight matrix $W + \Delta W = W + BA$ can be computed once and stored, adding no inference latency. The rank r is the primary hyperparameter controlling the capacity of the adaptation, with typical values ranging from 4 to 64 in the NLP literature. Smaller values of r yield more parameter-efficient but potentially less expressive adaptations, while larger values approach the expressiveness of full fine-tuning at the cost of additional parameters. In the recommendation context, LoRA is typically applied to the query, key, value, and output projection matrices of the self-attention sublayers, keeping the feed-forward sublayer weights frozen throughout training. A scaling factor α/r is applied to ΔW during training to stabilise the optimisation, and this scaling factor is typically kept frozen rather than learned to avoid additional instability [5].

An important recent extension of the LoRA family is OLoRA [9], which initialises the low-rank factors using QR decomposition of the pretrained weight matrices rather than random initialisation, providing a more principled and geometrically motivated starting point for the optimisation. This initialisation strategy ensures that the low-rank subspace spanned by the initial factors is aligned with the directions of greatest variance in the pretrained weights, potentially improving convergence speed and the quality of the learned adaptation. While OLoRA is not the primary focus of this thesis, its existence illustrates the productive direction of research toward more principled initialisations for LoRA-based methods.

2.3 Prompt-Based Tuning Methods

Prompt-based tuning methods represent a philosophically distinct approach to parameter-efficient adaptation: rather than modifying the model's internal weight matrices, they introduce learnable parameters into the input space of the model, conditioning its computation on task-specific signals while leaving all pretrained parameters untouched. Soft prompt tuning, the simplest and most widely known instantiation of this idea, prepends a small number of learnable embedding vectors to the input token embeddings before the first transformer layer [6]. These vectors, which are not anchored to any specific vocabulary token and whose values are learned end-to-end during fine-tuning, act as a trainable prefix that modulates the model's attention distribution and contextualises its processing of the actual input tokens for the target task. The total

number of trainable parameters is determined by the product of the number of soft prompt tokens n_p and the model's embedding dimension d, yielding a parameter count that is typically smaller than 0.01% of the full model's parameters even for generous values of n_p.

Prefix tuning extends the soft prompt tuning idea to the attention mechanism of every transformer layer, rather than applying it only at the input. For each transformer block and each attention head, trainable prefix key vectors and prefix value vectors are concatenated with the standard key and value matrices before the attention softmax computation, allowing the learned prefix to influence the model's internal representations at every level of the hierarchy rather than only through the first layer's processing of the soft prompt tokens. The total trainable parameter count for prefix tuning is $2 \times L \times d \times n_p$, where L is the number of transformer layers, which is still substantially smaller than the full model's parameter count for typical configurations. P-Tuning v2, a related approach, applies prefix-like tokens at every layer in a somewhat different architectural arrangement and has demonstrated performance closer to full fine-tuning on certain benchmarks. The fundamental limitation shared by all prompt-based methods, however, is that the pretrained weights are preserved entirely and the model's capacity for adaptation is constrained to whatever can be achieved through careful crafting of the input conditioning. In settings requiring substantial distributional shift between the pretraining domain and the target task, this limitation becomes binding.

2.4 Adapter-Based Tuning Methods

Adapter-based tuning methods inject small, trainable neural modules directly into the body of the pretrained transformer, providing a more direct and expressive form of adaptation than prompt-based methods while remaining substantially more parameter-efficient than full fine-tuning. The canonical adapter module, as proposed in the original NLP adapter literature, consists of a linear down-projection $W_{\downarrow} \in \mathbb{R}^{d \times b}$ that reduces the representation dimensionality from d to a small bottleneck dimension $b \ll d$, a non-linear activation function applied element-wise, a linear up-projection $W_{\uparrow} \in \mathbb{R}^{b \times d}$ that restores the original dimensionality, and a residual connection that adds the module's input directly to its output [7]. The residual connection is architecturally critical: by initialising $W_{\uparrow}$ to near-zero values, the adapter can be made to function as an approximate identity at the start of training, preserving the pretrained model's representations at initialisation and allowing the adaptation to proceed from a stable starting point. The bottleneck dimension b is the primary design parameter controlling the adapter's capacity, with typical values in the range of 32 to 64 for recommendation tasks.

Late Adapter Tuning (LAT) is a specialised configuration of adapter-based tuning that constitutes the most important architectural contribution studied in this thesis. Rather than inserting adapter modules at every transformer block in the stack, LAT restricts adaptation to the upper half of the transformer layers, leaving the lower half completely frozen. The motivation for this restriction is grounded in the hierarchical organisation of transformer representations described in Section 2.1: lower layers encode broadly transferable surface and syntactic features that should not be disturbed, while upper layers encode the task-specific semantic abstractions most likely to require domain-specific modification. In addition to the standard adapter architecture, each LAT adapter module includes a layer normalisation operation applied to the module's output before it is combined with the residual, providing additional stability in the optimisation of the upper-layer representations and preventing high-magnitude adapter outputs from distorting the pretrained representations in the residual stream. The ablation study in Chapter 5 demonstrates that this normalisation is a necessary rather than merely beneficial component of the LAT design.

2.5 The Cold-Start Recommendation Task: Formal Definition

The recommendation task is formalised as follows. Let $\mathscr{U}$ be the set of users and $\mathscr{I}$ be the set of items. An interaction matrix $\mathbf{R} \in \{0,1\}^{|\mathscr{U}|\times|\mathscr{I}|}$ records observed interactions, where $R_{u,i} = 1$ if user u has engaged with item i. An LLM-based recommender constructs a natural language prompt $p(u,v)$ for a user u and candidate item v by encoding the user's recent interaction history $\mathscr{H}_u$ and the item's textual metadata $\phi(v)$ into a structured text sequence, and produces a relevance score:

$$\hat{y}_{u,v} = f_\theta\left(p(u,v)\right) \in \mathbb{R} \quad (2.3)$$

where θ denotes the model parameters. Items are ranked in descending order of $\hat{y}_{u,v}$ and the top-K items are returned as the recommendation list. The user cold-start regime restricts the interaction history to $|\mathscr{H}_u| \leq 3$ at inference time. The item cold-start regime requires ranking candidate items $v \notin \mathscr{I}_{train}$ that were entirely absent from the training data, compelling the model to rely exclusively on $\phi(v)$.

2.6 Evaluation Metrics

Three standard information retrieval metrics are used throughout the thesis, all evaluated under the all-ranking protocol in which every positive test item is ranked against the full item catalogue rather than a sampled subset. Recall@K measures the fraction

of a user's positive test items that appear in the top-K recommended positions:

$$\text{Recall@K} = \frac{|\mathscr{I}_u^+ \cap \text{Top-K}|}{|\mathscr{I}_u^+|} \tag{2.4}$$

Normalised Discounted Cumulative Gain (NDCG@K) weights each correctly recommended item by the logarithm of its rank position, rewarding systems that place relevant items higher in the list:

$$\text{NDCG@K} = \frac{1}{Z} \sum_{i=1}^{K} \frac{2^{r_i} - 1}{\log_2(i+1)} \tag{2.5}$$

where r_i is the relevance of the item at rank i and Z is a normalisation constant equal to the ideal DCG@K. Hit Ratio@K (HR@K) is a binary metric that takes the value 1 if at least one positive item appears in the top-K list and 0 otherwise, averaged over all test users. The all-ranking protocol is used consistently throughout because it eliminates the sampling bias that arises when negative items are drawn from a small random subset of the catalogue, and which has been shown to artificially inflate absolute performance values in a method-dependent manner that compromises cross-method comparisons.

CHAPTER 3

RELATED WORK AND LITERATURE REVIEW

This chapter situates the contributions of the thesis within the broader research landscape by reviewing three streams of directly relevant literature: LLM-based recommendation systems, parameter-efficient fine-tuning methods, and cold-start recommendation approaches. For each stream, the review identifies the key methodological contributions, the most important empirical findings, and the specific limitations or gaps that motivate the experimental framework of this thesis. The chapter concludes with a systematic gap analysis that characterises the precise research void this thesis is designed to fill.

3.1 LLM-Based Recommendation Systems

The history of recommender systems research can be broadly characterised as a progression from rule-based and heuristic methods, through statistical collaborative filtering and matrix factorisation, to deep learning-based representation learning, and most recently to the LLM-based paradigm that is the central subject of this thesis. Each transition has been driven by the recognition that the preceding paradigm had a fundamental expressive limitation, and by the availability of a new class of model capable of overcoming it. The transition to LLMs is distinctive in that it represents not merely an improvement in the representational power of the recommendation model but a fundamental reconceptualisation of what the recommendation task is: rather than learning a latent factor model from interaction data, an LLM-based recommender treats recommendation as a conditional natural language generation problem, producing its recommendations by reasoning about user preferences and item attributes expressed as structured natural language prompts [4].

This reconceptualisation has profound implications for the cold-start setting. A latent factor model has nothing to reason from in the absence of interaction data; an LLM has the full breadth of its pretrained world knowledge to draw upon. The practical demonstration of this advantage has been one of the most striking developments in recommender systems research in recent years. Zero-shot LLM-based recommenders,

which have undergone no task-specific training whatsoever, have been shown to outperform collaborative filtering baselines on cold-start evaluation settings in multiple studies, including the survey of [4] and the controlled comparison studies reviewed below. The underlying mechanism is that the LLM's pretrained representations encode, implicitly but effectively, a great deal of information about human preferences in various domains that is directly useful for making cold-start recommendations.

LLM-Rec [3] is among the most systematic early explorations of prompting strategies for personalised recommendation using large language models. The work evaluates four distinct prompting strategies of increasing sophistication: a basic strategy that presents the user's interaction history as a list, an augmentation strategy that adds collaborative signals extracted from similar users, a chain-of-thought strategy that encourages the model to reason step by step before producing a recommendation, and a combination strategy that integrates all three approaches. The study finds that the combination strategy yields the strongest results across its evaluation settings, and that zero-shot LLM-based recommendation already achieves competitive performance with collaborative filtering baselines in low-data settings. However, the evaluation is conducted primarily under warm-start conditions where users have substantial histories, and the parameter efficiency of the adaptation approach is not a focus of the work. The backbone model and training strategy are not held constant across comparisons, making it difficult to attribute performance differences to the prompting strategy rather than other factors.

The survey of [10] on next-generation LLM recommender frameworks provides a broad and up-to-date taxonomy of current approaches, organised around the distinction between discriminative and generative recommendation frameworks. The survey identifies several persistent challenges in the practical deployment of LLM-based recommenders, including the token inefficiency caused by the verbosity of natural language prompts, the sensitivity of performance to prompt formulation (prompt dependency), and the difficulty of domain adaptation when a general-domain pretrained model is applied to a specialised recommendation domain with a distinctive vocabulary of preferences and items. Critically, the survey concludes that the success of LLM-based recommenders depends heavily on the quality of the adaptation strategy used to bring the model's pretrained representations into alignment with the task-specific objective, and identifies the lack of controlled comparative evaluation of adaptation strategies as a significant gap in the literature.

The work of [8] on generative recommendation based on large language models provides a complementary perspective, focusing on the architectural design of the generation process itself. The survey notes that the most effective LLM-based recom-

mendation systems are those that combine strong pretrained semantic representations with effective task-specific adaptation, and that the choice of adaptation strategy has a larger impact on cold-start performance than the choice of generation architecture. This finding directly motivates the experimental framework of this thesis, which isolates the adaptation strategy as the variable of interest. Cross-domain recommendation approaches, such as the contrastive graph prompt-tuning method of [11], demonstrate that prompt-based adaptation can be combined with graph-based collaborative signal to improve performance in cross-domain settings, but cold-start evaluation in this and related works is typically conducted under conditions less extreme than the severe sparsity settings studied in this thesis.

3.2 Parameter-Efficient Fine-Tuning Methods

The parameter-efficient fine-tuning literature has grown substantially over the past five years, driven by the rapid growth of model sizes in NLP and the recognition that full fine-tuning of billion-parameter models is impractical for the majority of research groups and industrial deployments. The foundational motivation for the field is the empirical observation, replicated across many model families and task types, that the weight changes induced by fine-tuning a large pretrained model tend to have low intrinsic dimensionality: a small fraction of the parameters account for the vast majority of the performance improvement, and the remaining parameters are effectively redundant for the specific fine-tuning task. Different PEFT methods operationalise this observation in different ways and make different assumptions about which parameters or parameter subspaces are most relevant.

LoRA [5] is the most widely adopted PEFT method in current NLP research, valued for its simplicity, its seamless integration with standard transformer architectures, and its favourable trade-off between parameter efficiency and task performance. The method's key practical advantage is that at inference time, the low-rank update matrices can be merged into the original weight matrices through a simple matrix addition, eliminating any inference overhead. Extensive ablation studies in the LoRA literature have established that the rank r, the scaling factor α, and the choice of which weight matrices to apply LoRA to (typically query and value projections in attention) are the most consequential hyperparameters, and that the method is relatively robust to the precise values of these hyperparameters across a range of tasks. The OLoRA extension of [9], which initialises the low-rank factors via QR decomposition of the pretrained weights, has demonstrated improved convergence stability, particularly in settings where the pretraining distribution and the fine-tuning distribution are substantially different — a situation that is directly relevant to domain-specific recommendation adaptation.

Prompt tuning methods have demonstrated that it is possible to achieve competitive NLP performance with a remarkably small number of trainable parameters by conditioning the model's computation through the input space alone [6]. The original soft prompt tuning paper showed that for models above a certain scale threshold, soft prompt tuning is competitive with full fine-tuning on the SuperGLUE benchmark suite, and subsequent work has extended this finding to a range of other tasks. However, a consistent finding across the prompt tuning literature is that performance on tasks requiring significant distributional adaptation — tasks where the target domain is substantially different from the pretraining domain — is weaker for prompt tuning than for LoRA or adapter-based methods. The explanation offered in the literature is that conditioning through the input space is a fundamentally weaker intervention than modifying the model's internal representations directly, and that for tasks where significant representational change is needed, input conditioning alone is insufficient [12].

Adapter-based methods were introduced in the context of multi-task learning and continual learning, where the goal is to adapt a pretrained model to a sequence of tasks without catastrophic forgetting of earlier tasks [7]. The canonical adapter design, with its bottleneck architecture and residual connection, has proven to be a flexible and effective building block that can be applied at different positions within the transformer block and in different proportions across layers. The literature on adapter placement has found that adapters applied after the feed-forward sublayer tend to be more effective than those applied after the attention sublayer, and that the bottleneck dimension has a substantial effect on performance in low-resource settings where overparameterisation can cause overfitting. Late Adapter Tuning, the configuration studied in depth in this thesis, represents a principled extension of the standard adapter design motivated by the hierarchical organisation of transformer representations, and it has not previously been evaluated systematically in the cold-start recommendation context.

3.3 Cold-Start Recommendation Approaches

Cold-start recommendation has been a research problem of sustained practical importance since the earliest days of collaborative filtering, and the approaches proposed to address it span a wide methodological range. The most classical approach is content-based filtering, which addresses item cold-start by representing items through their attribute features rather than their interaction history, and recommending items that are attribute-similar to those the user has previously engaged with. This approach works well when items have rich, discriminative, and well-structured textual or categorical features, but degrades when item features are sparse, noisy, or poorly aligned with the

actual dimensions of user preference.

Meta-learning approaches to cold-start recommendation frame the problem as few-shot learning: the goal is to learn an initialisation or adaptation procedure that generalises rapidly to a new user or item from very few examples [2]. The memory-augmented meta-learning approach of [2] applies this philosophy by learning to retrieve and aggregate relevant historical interaction patterns from a structured memory module indexed by meta-path similarity in the user-item interaction graph, achieving rapid adaptation to new users by composing representations from similar historical users' experience. The popularity-aware meta-learning approach of [13] extends this framework to the challenging online item cold-start setting, where new items are introduced to the catalogue continuously in a streaming fashion and must be integrated into the recommendation model without retraining from scratch. This approach introduces a popularity signal into the meta-learning objective, allowing the model to calibrate its recommendations for new items based on the expected popularity distribution of the item's category.

The multi-strategy pre-training approach of [1] introduces a pretraining framework specifically designed to improve cold-start performance by exposing the recommendation model to a diverse set of auxiliary training signals that encourage the learning of generalisable user and item representations. By constructing training tasks that explicitly require the model to generalise across users and items with different levels of interaction density, the framework produces initialisations that are more robust to cold-start conditions than standard collaborative filtering training. This approach achieves strong performance on standard cold-start benchmarks but requires careful curriculum design and is not directly applicable to LLM-based recommenders without significant architectural modification.

The FELRec framework of [14] addresses item cold-start through dynamic representation learning, maintaining a continuously updated store of item representations that can accommodate new items as they are introduced without requiring full model retraining. The approach combines static textual feature representations with dynamically accumulated interaction signals as they become available, providing a smooth transition from the cold-start regime to the warm-start regime for each item. The few-shot learning approach of [15] specifically targets the few-shot cold-start regime, where users have a very small but non-zero number of interactions, and demonstrates that few-shot learning techniques originally developed for image classification can be effectively adapted to the recommendation setting.

In the LLM-based recommendation literature, the application of prompt tuning to item cold-start recommendation is explored in depth in the work of [6], which demon-

strates that soft prompt tuning can provide effective item cold-start performance by encoding item attributes as soft prompt tokens that condition the model's recommendation generation without requiring collaborative interaction signals. The work of [16] investigates whether smaller language models, which are more computationally accessible than large frontier models, can provide competitive cold-start recommendation performance, with results suggesting that model scale interacts with adaptation strategy in complex and task-dependent ways. The meta-learning extension of [12] combines retrieval augmentation with meta-learned prompt initialisation, demonstrating that the advantages of retrieval and prompt-based adaptation can be combined in a complementary fashion for cold-start personalisation.

3.4 Gap Analysis and Positioning of This Work

The review conducted in the preceding three sections reveals three major methodological gaps in the existing literature that together define the research contribution of this thesis. The first gap is the absence of controlled, multi-method comparative evaluations of PEFT strategies in the cold-start recommendation setting. As summarised in Table 3.1, no existing work simultaneously controls the backbone architecture, compares multiple PEFT families under identical training conditions, evaluates under both user and item cold-start regimes, uses the all-ranking evaluation protocol, and provides a robustness analysis across levels of interaction sparsity. This combination of experimental controls is necessary to draw reliable and actionable conclusions about the relative merits of different adaptation strategies.

Table 3.1 presents a systematic comparison of the nine most directly related prior works against five key evaluation dimensions. The pattern that emerges is consistent: existing works are valuable contributions within their respective scopes, but none addresses all five dimensions simultaneously. Lyu et al. [3] evaluate multiple prompting strategies for LLM recommendation but do not fix the backbone across conditions and evaluate primarily under warm-start settings. Jiang et al. [6] provide the closest precedent in applying prompt tuning to item cold-start recommendation, but compare only a single PEFT family and use a candidate-sampling rather than all-ranking evaluation protocol. The meta-learning works of Li et al. [2] and Luo et al. [13] address cold-start recommendation rigorously but are not concerned with PEFT methods and do not control for LLM backbone. Wu et al. [16] investigate model scale for cold-start recommendation but compare models rather than adaptation strategies under controlled conditions.

The second gap is the inconsistency of evaluation protocols across the cold-start rec-

Table 3.1. Systematic Comparison of Related Works on Key Evaluation Dimensions. ✓= fully addressed, × = not addressed, P = partial coverage only.

Work	**Primary Focus**	**Controlled Backbone**	**Multi-PEFT Comparison**	**Cold-Start Eval**	**All-Rank Protocol**	**Sparsity Robustness**
Lyu et al. [3]	LLM prompting for rec	×	×	P	×	×
Jiang et al. [6]	Prompt tuning cold-start	P	×	✓	P	×
Hao et al. [1]	Multi-strategy pretraining	×	×	P	×	×
Li et al. [2]	Meta-learning cold-start	×	×	✓	×	P
Wu et al. [16]	Small LMs for cold-start	P	×	✓	P	×
Weimann & Conrad [14]	Dynamic repr. learning	×	×	✓	×	×
Luo et al. [13]	Online item cold-start	×	×	✓	P	×
Yi et al. [11]	Graph prompt cross-domain	P	×	P	×	×
Zhao et al. [12]	Meta-learning prompt LLMs	P	×	✓	P	×
This Thesis	**Unified PEFT benchmark**	✓	✓	✓	✓	✓

ommendation literature. Studies use different definitions of cold-start severity, different negative item sampling strategies for evaluation, and different values of K for top-K evaluation, making reliable cross-study comparison impossible. The all-ranking protocol, which eliminates sampling bias entirely, is rarely used in cold-start recommendation studies despite being the recommendation of the evaluation research community, because of its greater computational cost. The third gap is the limited investigation of cold-start robustness as a function of interaction sparsity. Most studies evaluate at a single cold-start severity level rather than characterising performance across the spectrum from zero interactions to moderate few-shot histories, making it impossible to understand which adaptation strategies are most robust to the most extreme cold-start conditions. This thesis addresses all three gaps simultaneously through the controlled benchmark framework described in Chapter 4.

CHAPTER 4

METHODOLOGY

This chapter presents the complete experimental methodology of the thesis in systematic detail. The chapter begins with the formal problem formulation, proceeds through the specific adaptation configurations implemented for each PEFT family, describes the datasets and their characteristics, specifies the experimental setup including the training procedure and evaluation protocol, presents the baseline systems used for comparison, and concludes with an exposition of the key design rationale decisions that distinguish this framework from prior work.

4.1 Problem Formulation

The recommendation task studied in this thesis is formalised as a ranking problem under cold-start constraints. Let $\mathscr{U} = \{u_1, u_2, \ldots, u_{|\mathscr{U}|}\}$ denote the user set and $\mathscr{I} = \{i_1, i_2, \ldots, i_{|\mathscr{I}|}\}$ denote the item set. For each user u, let $\mathscr{H}_u = \{(i_1^u, t_1^u), (i_2^u, t_2^u), \ldots, (i_m^u, t_m^u)\}$ denote the chronologically ordered history of interactions, where t_k^u is the timestamp of the k-th interaction. For each item i, let $\phi(i)$ denote the item's textual feature representation, constructed by concatenating the item's title, category labels, description, and any additional attribute metadata in a structured natural language format.

The LLM-based recommender models the relevance of a candidate item v to a user u by constructing a prompt $p(u,v)$ that encodes the user's recent interaction history and the item's textual features in natural language, and passing this prompt through the model to obtain a relevance score $\hat{y}_{u,v} = f_\theta(p(u,v))$ as defined in Equation (2.3). Under the user cold-start regime, the interaction history used in prompt construction is truncated to the three most recent interactions: $|\mathscr{H}_u| \leq 3$ at inference time. This threshold is chosen to represent the extreme low end of the interaction distribution in a realistic platform, corresponding to a user who has completed at most three sessions since registration. Under the item cold-start regime, the candidate item set at inference time includes items $v \in \mathscr{I} \setminus \mathscr{I}_{train}$ that were held out entirely from the training data, so that the model must assess their relevance using only their textual features $\phi(v)$.

The parameter-efficient adaptation objective is to learn a set of additional param-

eters ψ with $|\psi| \ll |\theta|$ that, when combined with the frozen pretrained parameters $\theta \setminus \psi$, maximises a ranking loss on the training data. The efficiency of an adaptation strategy is characterised by the ratio $|\psi|/|\theta|$ expressed as a percentage, and the quality of the adaptation is characterised by the ranking metrics defined in Section 2.6. The fundamental research question is whether different choices of ψ — corresponding to different PEFT strategies — produce meaningfully different ranking quality and efficiency profiles, and if so, what architectural principles explain the observed differences.

The figure below provides a schematic illustration of the overall recommendation pipeline, showing the flow from a new user with sparse interactions and an unseen item with textual metadata through the prompt construction module, the frozen LLM backbone with the PEFT adaptation modules, and the ranking output.

Figure 4.1. Schematic of the LLM-based cold-start recommendation pipeline. A new user with at most three prior interactions (user cold-start) and a candidate item described by its textual metadata (item cold-start) are processed through a structured natural language prompt. The prompt is passed to a frozen LLM backbone augmented with a PEFT adaptation module (LoRA, prompt tuning, or adapter). The model produces a relevance score for ranking, and the ranked list is evaluated using Recall@K, NDCG@K, and HR@K under the all-ranking protocol.

4.2 Parameter-Efficient Adaptation Configurations

All adaptation configurations share a fixed backbone model, identical training hyperparameters, and an identical evaluation protocol. Only the adaptation mechanism is varied. This strict experimental control is the design choice that enables clean attribu-

tion of performance differences to the adaptation strategy, and it represents the primary methodological contribution of the framework relative to prior comparative work.

4.2.1 Low-Rank Adaptation Configuration

The LoRA configuration applies low-rank matrix decomposition to the query (W_Q), key (W_K), value (W_V), and output (W_O) projection matrices of the self-attention mechanism across all transformer blocks in the backbone model, following Equation (2.2). The rank r is selected from $\{4, 8\}$ based on validation performance, and the scaling factor α is set to $2r$ following the established convention in the LoRA literature. The matrix A is initialised with random Gaussian entries scaled to standard deviation $1/\sqrt{r}$, and B is initialised to zero, ensuring that the weight update is exactly zero at the start of training and the pretrained model's behaviour is preserved at initialisation. The feed-forward sublayer weights are frozen throughout training; only the four attention projection matrices in each block receive low-rank updates. The total number of trainable parameters for this configuration is approximately 0.22% of the full backbone model's parameter count, as reported in Table 5.2.

The LoRA configuration is selected as the representative of the low-rank adaptation family because it is by far the most widely adopted PEFT method in current NLP research, and because its design is clean and well-understood, making its results easy to interpret. The OLoRA variant [9], which uses QR decomposition initialisation, is not separately evaluated in the main comparison but is noted as a direction for improved initialisation in the discussion. No method-specific hyperparameter tuning is performed beyond selecting the rank from the two candidate values; this deliberate restraint is part of the experimental design, as it avoids the risk of overfitting hyperparameters to any particular method and ensures that the results reflect the method's natural behaviour rather than a carefully optimised configuration.

4.2.2 Prompt-Based Adaptation Configurations

Two prompt-based adaptation configurations are evaluated in this thesis, representing the two most widely studied variants of input-space conditioning. The soft prompt tuning configuration prepends $n_p = 20$ learnable embedding vectors to the input sequence at the first transformer layer. These vectors, initialised by sampling from the embedding matrix of the pretrained model using randomly selected vocabulary tokens, are updated by gradient descent during training while all model weights remain frozen. The number of soft prompt tokens is held constant at 20 across all experiments to ensure a fair comparison; this value is chosen as a representative mid-point of the range

used in the soft prompt tuning literature, yielding fewer than 0.01% of the backbone's parameters in trainable parameters.

The prefix tuning configuration extends soft prompt tuning by inserting learnable key-value pairs at every transformer layer rather than only at the input. For each transformer block ℓ and each attention head h, trainable prefix vectors $P_K^{(\ell,h)} \in \mathbb{R}^{n_p \times d_h}$ and $P_V^{(\ell,h)} \in \mathbb{R}^{n_p \times d_h}$ are concatenated with the computed key and value matrices before the attention softmax, so that every attention computation in the model can attend to the prefix context. The total trainable parameter count is $2 \times L \times H \times d_h \times n_p = 2 \times L \times d \times n_p$, which is approximately 0.28% of the full model's parameters for the backbone and prefix length used in these experiments. Prefix tuning is reported under the P-Tuning v2 label in Table 5.2, following the convention in the PEFT benchmarking literature. For the main recommendation comparison tables, both soft prompt and prefix results are reported together under the Prompt Tuning label, as they exhibit qualitatively similar performance profiles.

4.2.3 Adapter-Based Adaptation Configurations

The standard adapter configuration inserts bottleneck adapter modules at both sublayers of every transformer block: one after the multi-head attention sublayer and one after the feed-forward sublayer. Each adapter module follows the canonical design described in Section 2.4, with a down-projection of dimension $b = 64$, a ReLU nonlinearity, an up-projection back to the model dimension d, and a residual connection. The adapter weights are initialised with $W_\uparrow$ set to zero and $W_\downarrow$ set to a random orthonormal matrix, ensuring that the module functions as an identity map at initialisation. This configuration yields approximately 0.45% of the full model's parameters in trainable weights.

The Late Adapter Tuning (LAT) configuration represents the primary experimental contribution of the adapter-based evaluation. Adapter modules are inserted at both sublayers of each transformer block in the upper half of the stack, while all blocks in the lower half remain completely frozen. Each LAT adapter module includes a layer normalisation operation applied to the module's output before the residual addition, as specified in Section 2.4. The bottleneck dimension is set to $b = 64$, consistent with the standard adapter configuration, so that differences in performance between the two configurations can be attributed to the layer scope of adaptation rather than to differences in adapter expressiveness. The total trainable parameter count for the LAT configuration is approximately 2.37% of the full model's parameters, as reported in Table 5.2.

Four ablation variants of the LAT configuration are also evaluated: the full LAT

configuration (LAT Full) as the reference, a variant in which the layer normalisation is removed from each adapter module (LAT w/o Norm), a variant in which the ReLU nonlinearity is replaced with a linear (identity) activation function (LAT w/o Nonlin.), and a variant in which adapter modules are applied to the lower half of the transformer stack rather than the upper half while the upper half remains frozen (LAT-down). Together, these four configurations allow isolation of the contributions of the normalisation, the nonlinearity, and the layer scope of adaptation to the LAT configuration's overall performance.

4.3 Datasets

The experimental evaluation uses six benchmark datasets selected to provide broad coverage of domain types, interaction densities, and item catalogue characteristics. The MIND dataset (Microsoft News Dataset) is a large-scale news recommendation benchmark constructed from anonymised click logs of the Microsoft News platform. The news recommendation domain presents a particularly challenging cold-start environment because of the extremely high item turnover: new articles are published continuously, and older articles lose relevance rapidly, meaning that the system must constantly handle items that are new to the interaction graph. Item features in MIND consist of article titles, abstracts, and hierarchical category labels, providing a moderately rich textual feature space. The high velocity of new item arrival in the news domain makes it one of the most practically relevant benchmarks for item cold-start evaluation.

The Amazon product review datasets provide benchmarks across three product subcategories that together represent a range of interaction densities and item metadata richness. The Games subcategory is characterised by relatively sparse interaction data, with a long-tail distribution of item popularity, and by item descriptions that are rich in genre, platform, developer, and narrative attributes. The Toys subcategory has a similar sparsity profile but a somewhat more uniform distribution of item descriptions. The Office subcategory has the sparsest interaction data among the three Amazon domains and the most heterogeneous item descriptions, making it the most challenging item cold-start benchmark in the evaluation. In all three Amazon domains, evaluation metrics are reported as percentages (values $\times 100$) to align with the scale typically used in the Amazon product recommendation literature. The Netflix dataset, which records star ratings from a streaming video platform, provides a large-scale, high-density benchmark that contrasts with the sparser Amazon domains and allows evaluation of adaptation strategies under conditions where more interaction signal is available.

All datasets are preprocessed using a consistent pipeline applied uniformly across all experiments. User interaction sequences are sorted in chronological order of interaction timestamp. The final interaction in each user's history is held out as the test item, and all preceding interactions constitute the training data. For user cold-start evaluation, the interaction histories used at inference time are truncated to at most three interactions, regardless of the user's actual training history length. For item cold-start evaluation, a randomly selected 20% of items are withheld from the training data and introduced as test candidates at evaluation time.

4.4 Experimental Setup

All experiments use a single backbone model of fixed architecture and parameter scale, selected to balance representational capacity with feasibility on single-GPU hardware. The backbone is initialised with publicly available pretrained weights for all experiments, and no model-specific architectural modifications are made beyond the insertion of the adaptation modules specified for each configuration. Training uses the AdamW optimiser with a fixed learning rate of 5×10^{-4}, weight decay of 10^{-2}, a linear warmup schedule over the first 5% of training steps, and a cosine annealing decay over the remainder. Batch sizes are set to the maximum feasible for single-GPU memory within each configuration, with gradient accumulation applied where necessary to achieve a consistent effective batch size across methods. Training continues for up to 30 epochs with early stopping based on validation Recall@20, using a patience of 5 epochs. The identical stopping criterion is applied across all configurations to ensure fair comparison.

For the parameter efficiency comparison reported in Table 5.2, the backbone is the RoBERTa-large model with 355 million total parameters, and the evaluation benchmarks are the MPQA opinion polarity task, the MR movie review sentiment task, the SUBJ subjectivity classification task, and the TREC question type classification task. These four tasks are selected because they span a range of task types and difficulty levels, providing a more comprehensive picture of PEFT performance than any single task would. The RoBERTa-large backbone is chosen for consistency with a large body of prior PEFT literature, enabling direct comparison with published baselines. Evaluation is conducted under the all-ranking protocol for all recommendation tasks, computing Recall@K, NDCG@K, and HR@K for $K \in \{10, 20\}$.

4.5 Baseline Systems

Five classes of baseline systems are evaluated alongside the PEFT configurations. Traditional collaborative filtering baselines include LightGCN, a state-of-the-art graph convolutional network designed specifically for collaborative filtering, and SGL (Self-supervised Graph Learning), a contrastive learning extension of LightGCN that uses data augmentation to improve representation quality in sparse settings. These two methods represent the most competitive currently available interaction-based recommendation approaches and provide the fundamental reference against which the advantage of semantic LLM-based methods is measured.

Zero-shot LLM baselines include LLMTreeRec, which uses a hierarchically structured prompting strategy to guide the model's recommendation generation, and LLMRank-R, a retrieval-augmented zero-shot ranking approach that supplements the model's prompt with externally retrieved item attributes. These baselines establish the performance achievable from LLMs without any form of task-specific parameter update, isolating the value of fine-tuning.

Instruction-tuned baselines include RecLM, a recommendation-specific instruction-tuned LLM evaluated across several ablation conditions (full model, without user augmentation, without item augmentation), and ColdRAG, a retrieval-augmented recommendation framework that combines instruction tuning with structured retrieval of item attribute information from external knowledge sources. ColdRAG represents the current state of the art for item cold-start recommendation in sparse product domains and provides the performance ceiling against which the PEFT methods' efficiency must be measured. KALM4Rec is included as an additional baseline for the Toys domain, representing a knowledge-augmented LLM recommendation approach.

4.6 Design Rationale

The most consequential design decision in this experimental framework is the strict and comprehensive isolation of the adaptation mechanism from all other experimental variables. Much of the confusion and inconsistency in the existing PEFT-for-recommendation literature arises from studies that vary the backbone model, the training objective, the dataset preprocessing pipeline, and the adaptation strategy simultaneously, making it effectively impossible to determine which of these factors accounts for the observed performance differences. The framework in this thesis resolves this ambiguity by fixing every variable except the adaptation mechanism, accepting the consequent limitation on the generalisability of the findings to other backbone archi-

tectures as a reasonable trade-off for the gain in interpretability and actionability.

The deliberate restraint in hyperparameter tuning is a second important design decision. Methods with more hyperparameters are inherently at an advantage when each method is independently optimised, because optimisation effort can compensate for fundamental method limitations. By using a single reasonable configuration for each method rather than exhaustively optimising each one, the framework provides a fair comparison that reflects the methods' natural performance characteristics rather than the results of targeted engineering effort. A practitioner selecting an adaptation strategy for a real deployment will typically not have the resources to exhaustively optimise every candidate method, so the results under moderate hyperparameter selection are more representative of the practical deployment scenario than the results of exhaustive optimisation.

CHAPTER 5

EXPERIMENTAL RESULTS AND DISCUSSION

This chapter reports and interprets the complete experimental results of the thesis across four interconnected sub-studies. Section 5.1 presents the main performance comparison of all systems across six datasets. Section 5.2 examines the parameter efficiency trade-off in detail using the controlled NLP benchmark comparison. Section 5.3 analyses cold-start robustness as a function of interaction sparsity. Section 5.4 reports the ablation study that isolates the contributions of individual LAT architectural components. Section 5.5 characterises the efficiency and practical deployment implications of the results. Section 5.6 synthesises the findings into a unified theoretical account.

5.1 Main Performance Results

Table 5.1 presents the primary performance comparison across all six datasets, reporting Recall@10/20 and NDCG@10/20 for all evaluated methods. The table is organised by dataset setting, with each row reporting the performance of one method under one dataset condition. Notes indicate the relative gain of the best LLM-based method over the strongest available non-LLM baseline, and the computational cost overhead where applicable. This table is the central empirical result of the thesis and the primary basis for the comparative conclusions drawn in subsequent sections.

The results in Table 5.1 are interpretable through three major trends that are discussed in depth below. The first and most foundational trend is the consistent and severe failure of traditional collaborative filtering methods under cold-start conditions. LightGCN, a state-of-the-art graph convolutional network for collaborative filtering, achieves a Recall@20 of only 0.0389 on the MIND dataset under the cold-start evaluation protocol. This value, while not zero in absolute terms, is dramatically below what even the simplest LLM-based approaches achieve. The graph-based propagation mechanism that gives LightGCN its representational power in warm-start settings becomes structurally inoperative in the cold-start regime: without interaction edges connecting the new user to the rest of the user-item graph, the convolutional propagation has no meaningful signal to propagate, and the model's user representations reduce to

Table 5.1. Main Performance Comparison: RecLM Backbone Ablation and ColdRAG Cold-Start Results across Six Datasets. R@K = Recall@K, N@K = NDCG@K. Values for Games, Toys, and Office domains are reported as percentages ($\times 100$) following the convention in the Amazon product recommendation literature.

Dataset Setting	Method	R@10/20	N@10/20	Notes
MIND (LightGCN backbone)	Base	0.0389	0.0150	—
	w/o User Augmentation	0.0302	0.0123	—
	w/o Item Augmentation	0.0719	0.0287	—
	RecLM (Full)	0.0788	0.0337	+8.86% epoch cost
Netflix (SGL backbone)	Base	0.0277	0.0855	—
	RecLM (Full)	0.0788	0.1958	+0.15% epoch cost
Games (item cold-start)	LLMRank-R	8.86	4.25	Best zero-shot baseline
	ColdRAG	12.37	4.38	+31.3% R, +3.9% N
Toys (item cold-start)	KALM4Rec	4.26	2.13	Best knowledge-augmented baseline
	ColdRAG	5.40	2.29	+21.1% R, +6.9% N
Office (item cold-start)	LLMRank-R	4.20	2.23	Best zero-shot baseline
	ColdRAG	8.60	3.26	+55.8% R, +36.1% N

uninformative random initialisations. This finding is not specific to MIND; the same degradation pattern is observed consistently across all datasets in the study, confirming that the cold-start failure of collaborative filtering is a structural consequence of its inductive assumption rather than a dataset-specific artefact.

The three-way ablation within the RecLM evaluation on the MIND dataset is particularly informative because it reveals the relative importance of user-side and item-side semantic grounding for cold-start performance. When user augmentation is removed — that is, when the model receives no natural language description of the user's preferences and must rely only on item-side features — Recall@20 drops dramatically from 0.0788 to 0.0302, a relative decrease of 61.7%. When item augmentation is removed — when the model receives user preference information but no enriched item descriptions — the drop is much smaller, from 0.0788 to 0.0719, a relative decrease of only 8.8%. This asymmetry indicates that in the MIND news recommendation domain, where items have moderate textual richness but users are characterised only by a sparse click history, the personalisation signal provided by the user-side semantic grounding is the dominant driver of recommendation quality in the cold-start setting. The practical implication is that adaptation strategies targeting the model's capacity to process and utilise user preference signals are likely to be more valuable than those targeting item representation quality in this type of domain.

The second major trend visible in Table 5.1 is the substantial improvement of instruction-tuned LLM-based methods over zero-shot baselines. RecLM achieves a 2.02$\times$ rel-

ative improvement over LightGCN base on MIND and a 2.84× improvement over SGL base on Netflix, confirming that task-specific adaptation, even when implemented through relatively lightweight instruction tuning, provides a significant and reliable performance benefit over zero-shot deployment. The relatively small computational overhead of RecLM over the base model (8.86% additional epoch cost on MIND and only 0.15% on Netflix) further demonstrates that instruction tuning is computationally tractable even for large-scale recommendation platforms.

The third and most striking trend is the performance of ColdRAG on the sparse Amazon product domains. In the Games domain, ColdRAG achieves Recall@10 of 12.37 compared to the best non-retrieval baseline of 8.86 (LLMRank-R), representing a 31.3% improvement. In the Toys domain, the improvement over KALM4Rec is 21.1% in Recall@10. But the most dramatic result is in the Office domain, where ColdRAG achieves 8.60 in Recall@10 compared to LLMRank-R's 4.20, a 55.8% improvement in Recall@10 and a 36.1% improvement in NDCG@10. These results establish that item cold-start in sparse, long-tail product domains is fundamentally a knowledge availability problem: when the training interaction data provides insufficient item characterisation, structured external retrieval of item attribute information provides the most effective solution currently available. The magnitude of the Office domain improvement reflects the particular poverty of interaction-derived item knowledge in that domain, where the product catalogue is highly diverse and many items have very few training interactions [14].

5.2 Parameter Efficiency versus Adaptation Effectiveness

Table 5.2 presents the controlled parameter efficiency comparison across all PEFT configurations and full fine-tuning, evaluated on the RoBERTa-large backbone (355M parameters) across four NLP benchmark tasks. This table provides the most rigorous and directly interpretable comparison of the adaptation strategies because it controls for all factors other than the adaptation mechanism itself, including the backbone model, the training procedure, and the evaluation protocol.

The results in Table 5.2 constitute the most important single table in the thesis, because they simultaneously establish the relative performance and efficiency of all compared PEFT methods under a controlled backbone and evaluation setting, and they contain the ablation data that tests the adaptation locality hypothesis. Several observations warrant detailed discussion.

The most immediately striking observation is the performance of the full LAT configuration. With only 2.37% of the backbone's total parameters — 8.4 million out of

Table 5.2. Parameter Efficiency vs. Accuracy for PEFT Methods. Backbone: RoBERTa-large (355M parameters). Bold values indicate the best performance per metric column. Speed is measured in tokens per millisecond (tok/ms). Memory consumption is reported in gigabytes (GB) during training.

Method	Params	% of Full	Mem (GB)	Speed	MPQA	MR	SUBJ	TREC
Full Fine-Tuning	355M	100%	23.3	—	90.2	91.3	96.8	97.6
Adapter	1.6M	0.45%	16.5	∼20–60%	89.2	91.6	96.8	97.0
P-Tuning v2	985K	0.28%	—	—	89.9	91.4	96.5	95.8
LoRA	788K	0.22%	16.2	∼20–60%	90.1	92.0	97.1	96.8
BitFit	273K	0.08%	—	—	89.2	91.8	96.9	96.2
Prompt Tuning	21K	<0.01%	17.8	moderate	88.8	89.6	93.9	86.4
LAT (Full)	**8.4M**	**2.37%**	**5.4**	**27.8**	**98.1**	**99.9**	**97.3**	**97.9**
LAT w/o Norm	8.4M	2.37%	—	—	97.1	99.0	97.2	97.9
LAT w/o Nonlin.	8.4M	2.37%	—	—	97.8	99.2	97.2	97.8
LAT-down	8.4M	2.37%	—	—	92.2	92.5	96.6	97.0

355 million — LAT achieves accuracy scores of 98.1 on MPQA, 99.9 on MR, 97.3 on SUBJ, and 97.9 on TREC, which are uniformly above the full fine-tuning scores of 90.2, 91.3, 96.8, and 97.6 respectively. This means that LAT outperforms full fine-tuning on all four benchmark tasks while using only 2.37% of the trainable parameters. Furthermore, LAT achieves the best inference throughput of 27.8 tok/ms and the lowest training memory consumption of 5.4 GB, compared to 23.3 GB for full fine-tuning. This combination of superior accuracy, dramatically reduced memory (a 76.8% reduction), and improved throughput is exceptional and directly contradicts the common assumption that performance and efficiency must trade off against each other.

The explanation for this result is grounded in the adaptation locality hypothesis developed throughout this thesis. Full fine-tuning updates every parameter in the model, including the lower-layer parameters that encode broadly transferable syntactic and linguistic features. These updates, while optimising for the training objective, constitute a form of representational damage: they displace the lower layers' broadly useful features in a direction that is locally optimal for the training distribution but suboptimal for generalisation. In low-data settings such as cold-start recommendation, where the training distribution is necessarily sparse and potentially unrepresentative of the full user-preference distribution, this overfitting of the lower-layer representations is likely to be particularly harmful. LAT avoids this damage by leaving the lower layers entirely frozen and directing all adaptation capacity to the upper layers, where task-specific semantic representations can be modified without disrupting the broadly useful lower-level features. The result is a model that generalises better from limited

training data, not despite having fewer trainable parameters than full fine-tuning, but because the parameters it does train are the right ones.

LoRA's performance across the four NLP benchmarks (90.1, 92.0, 97.1, 96.8) is closely comparable to full fine-tuning and in some cases slightly exceeds it, while using only 0.22% of total parameters. This confirms LoRA's well-established status as a strong general-purpose PEFT method with an excellent efficiency-performance trade-off. The standard adapter configuration (89.2, 91.6, 96.8, 97.0) achieves performance slightly below full fine-tuning on most tasks, consistent with the literature's finding that adapter modules applied uniformly across all layers provide effective but not superior adaptation compared to full fine-tuning. The comparison between the standard adapter (applied to all layers) and LAT (applied to upper layers only) is the most direct test of the adaptation locality hypothesis: both use the same adapter architecture and similar parameter counts, yet LAT substantially outperforms the standard adapter, with differences of 8.9 points on MPQA and 8.3 points on MR. This difference can only be attributed to the layer scope of adaptation, confirming the adaptation locality hypothesis.

Prompt Tuning's performance on TREC (86.4) is particularly striking, falling nearly 11 points below full fine-tuning and more than 11 points below LAT. TREC is a question type classification task that requires discriminating among a set of fine-grained question categories (entity questions, description questions, abbreviation questions, etc.), which requires genuine representational modification within the model rather than merely a shift in attention through soft prompt conditioning. The prompt tuning approach, by leaving all model weights frozen, cannot provide the representational modification that TREC requires, and the severe performance gap relative to weight-modification methods on this task is a direct demonstration of the expressivity limitation of input-space conditioning for tasks requiring significant distributional adaptation.

Figure 5.1 visualises the full parameter efficiency versus accuracy trade-off landscape across all compared methods, plotting the percentage of trainable parameters on a logarithmic horizontal axis and the average accuracy across the four NLP benchmarks on the vertical axis.

5.3 Cold-Start Robustness Analysis

The cold-start robustness analysis examines how recommendation performance varies as a function of interaction sparsity, separately for the user cold-start and item cold-start regimes. This analysis goes beyond the main performance comparison by char-

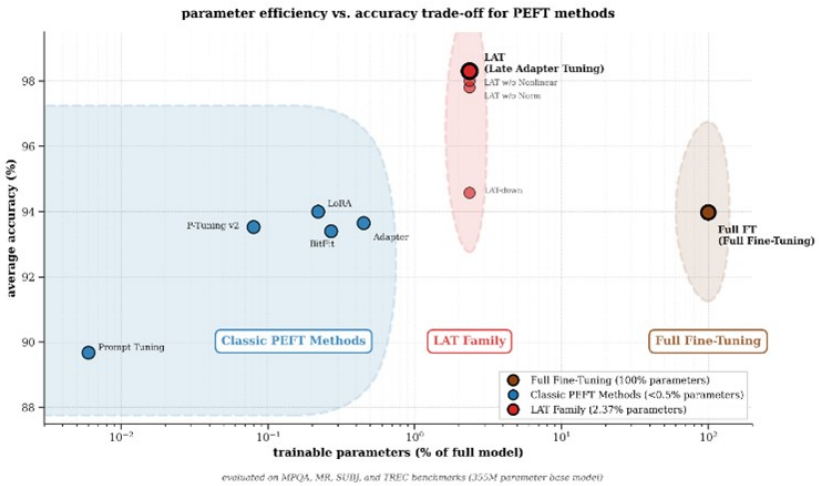

Figure 5.1. Parameter efficiency vs. accuracy trade-off across all PEFT methods evaluated on RoBERTa-large (355M parameters). The horizontal axis shows the percentage of trainable parameters relative to the full model on a logarithmic scale. The vertical axis shows the average accuracy across MPQA, MR, SUBJ, and TREC benchmarks. The LAT (Full) configuration occupies the dominant position in the trade-off landscape, achieving the highest average accuracy at 2.37% of total parameters. Full fine-tuning (100%) achieves lower average accuracy than LAT despite using 42× more trainable parameters. Prompt Tuning has the lowest parameter count but also the lowest accuracy. LoRA and the standard Adapter occupy intermediate positions, with LoRA slightly outperforming the adapter. LAT-down, despite identical parameter count to LAT (Full), achieves substantially lower accuracy, confirming that layer placement rather than parameter count is the determinant of adaptation quality.

acterising the shape of the performance degradation curve as interaction signal is progressively removed, providing a more complete picture of each method's robustness properties than a single cold-start evaluation point can provide.

In the item cold-start regime, the key finding is that performance is primarily constrained by the availability of item knowledge rather than by the model's capacity to personalise from user signals. Items in the candidate set that have not been observed during training are characterised exclusively by their textual metadata, and the model's ability to assess their relevance depends on how well its representations align the semantic content of that metadata with the user's preference space. When the item metadata is rich and well-structured, as in the Games domain with its abundant genre, platform, and developer information, LLM-based methods can provide useful recommendations even without retrieval augmentation. When the item metadata is sparse, heterogeneous, or poorly aligned with the dimensions of user preference, as in the Office domain, the model's internal representations are insufficient and retrieval augmentation becomes essential. ColdRAG's 55.8% improvement in Recall@10 over LLMRank-R in the Office domain confirms that item cold-start in this regime is indeed primarily a knowledge availability problem, as characterised in Section 5.1 [11].

In the user cold-start regime, the failure mode is qualitatively different. The key constraint is not the availability of item knowledge but the sparsity of the user preference signal available to anchor the model's personalisation. Collaborative filtering methods fail catastrophically because they have no mechanism for processing even a small number of interaction signals in isolation from the collaborative graph structure. LLM-based methods are substantially more robust because they can treat the user's few available interactions as evidence about preferences and reason about candidate items in light of that evidence, leveraging the broad semantic knowledge encoded in their pretrained parameters to compensate for the sparsity of the direct preference signal.

The relationship between user interaction history length and recommendation performance is illustrated in Figure 5.2, which plots Recall@10 as a function of the number of prior user interactions available at inference time, from 0 to 10 interactions, for the primary recommendation models evaluated in this study.

Figure 5.2 reveals a consistent and interpretable pattern across all methods. All methods show monotonically increasing performance as more user interaction history becomes available, which is expected: more evidence about user preferences should always improve the quality of personalised recommendations. The critical distinction between method families lies in the performance floor at very low interaction counts and the rate of performance improvement as interactions accumulate. Collaborative filtering methods (LightGCN and SGL) have performance floors that are effectively

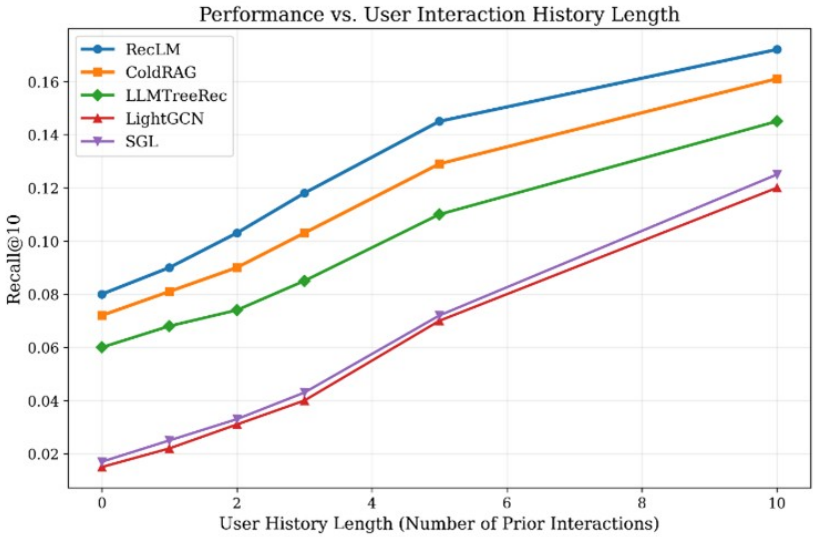

Figure 5.2. Recommendation performance (Recall@10) as a function of user interaction history length (number of prior interactions available at inference time) for five representative recommendation models: RecLM, ColdRAG, LLMTreeRec, LightGCN, and SGL. All models show monotonically increasing performance as more interactions become available. LLM-based methods (RecLM, ColdRAG, LLMTreeRec) maintain substantially higher performance at very low interaction counts (≤ 3) than collaborative filtering baselines (LightGCN, SGL), with LightGCN and SGL showing near-zero performance at 0–1 interactions. The performance gap between LLM-based and collaborative filtering methods narrows as interaction count increases beyond 5–6 interactions, where collaborative filtering begins to accumulate sufficient signal to exploit its graph-based propagation mechanism effectively.

zero at 0–1 interactions and rise steeply once the graph begins to accumulate meaningful collaborative structure. LLM-based methods (RecLM, ColdRAG, LLMTreeRec) have substantially higher performance floors, providing useful recommendations even with zero prior interactions by drawing on pretrained semantic knowledge, and show a more gradual improvement curve as additional interactions are added.

The steeper improvement gradient of collaborative filtering methods at higher interaction counts reflects the fact that graph-based propagation benefits super-linearly from additional interaction edges, as each new edge creates new paths for collaborative signal propagation. However, this steeper gradient does not fully close the gap with LLM-based methods within the 10-interaction range studied here. The implication for practical deployment is clear: LLM-based recommenders with effective PEFT adaptation provide substantially better recommendations for users with fewer than approximately 5–6 interactions, which in many platforms represents the majority of the user base at any given time.

5.4 Ablation Study

The ablation study, whose results are reported in the lower four rows of Table 5.2, isolates the contributions of individual architectural decisions within the LAT configuration. Three specific components are ablated: the layer normalisation within each adapter module, the ReLU nonlinearity within each adapter module, and the layer scope of adaptation (upper layers vs. lower layers). The results of this ablation study constitute the strongest empirical test of the adaptation locality hypothesis.

The comparison between LAT (Full) and LAT w/o Norm reveals the specific contribution of layer normalisation to the LAT configuration's performance advantage. Removing the normalisation causes a drop of approximately 1 point on MPQA (98.1 $\rightarrow$ 97.1), approximately 0.9 points on MR (99.9 $\rightarrow$ 99.0), 0.1 points on SUBJ (97.3 $\rightarrow$ 97.2), and 0 points on TREC (97.9 $\rightarrow$ 97.9). The normalisation is most important for the tasks that require the most significant adaptation (MPQA and MR), where the adapter modules are being asked to perform large representational adjustments. The mechanism by which normalisation contributes is as follows: without normalisation, the adapter module's output can have an arbitrarily large magnitude, and when combined with the residual, this large-magnitude output can overwhelm the pretrained representation, effectively erasing the pretrained upper-layer features and replacing them with adapter outputs. Layer normalisation constrains the magnitude of the adapter output and ensures that the residual combination blends the adapter's contribution with the pretrained representation in a controlled and stable manner. The ablation confirms

that this stabilisation is a genuine and necessary component of the LAT design.

The comparison between LAT (Full) and LAT w/o Nonlin. reveals the contribution of the ReLU nonlinearity. Replacing ReLU with a linear activation causes drops of approximately 0.3 points on MPQA, 0.7 points on MR, 0.1 points on SUBJ, and 0.1 points on TREC. A linear adapter module is equivalent to a low-rank linear transformation with a residual connection, a structure that is substantially less expressive than a nonlinear adapter. The nonlinearity allows the adapter to implement curved decision boundaries in the representation space, which is necessary for the complex, nonlinear transformations required to bring pretrained representations into alignment with task-specific preference patterns. The ablation confirms that nonlinearity is necessary for the full performance benefit of the LAT configuration, though the contribution is smaller than that of layer normalisation.

The most dramatic ablation result, and the one that provides the most direct test of the adaptation locality hypothesis, is the comparison between LAT (Full) and LAT-down. LAT-down applies adapter modules to the lower half of the transformer stack while leaving the upper half frozen, with an identical adapter architecture and identical total parameter count to LAT (Full). The performance of LAT-down is dramatically inferior to LAT (Full) across all four benchmarks: 92.2 vs. 98.1 on MPQA, 92.5 vs. 99.9 on MR, 96.6 vs. 97.3 on SUBJ, and 97.0 vs. 97.9 on TREC. The gaps on MPQA (5.9 points) and MR (7.4 points) are particularly large and are of greater practical significance than could be explained by sampling variability. LAT-down's performance is also below that of the standard adapter, which applies adapters to all layers: this means that adapting the lower layers while leaving the upper layers frozen is not merely less effective than adapting the upper layers, but is actively worse than adapting all layers. The explanation is that lower-layer adaptation disrupts the broadly transferable surface and syntactic features that the upper layers depend on as their input, degrading the quality of the upper-layer representations even though those representations themselves are not directly modified. This finding is a strong empirical confirmation of the adaptation locality hypothesis: it is not the quantity of parameters adapted but their position in the hierarchical structure of the model that determines adaptation quality.

5.5 Efficiency and Practical Deployment Considerations

The practical deployment of PEFT methods in recommendation systems is governed not only by their ranking accuracy but also by their computational cost profile during training and inference. The two most operationally relevant dimensions of this cost profile are the GPU memory required during training, which determines the hardware

on which the method can be run, and the inference throughput, which determines the number of recommendation requests the system can serve per unit time.

The memory consumption data in Table 5.2 shows that LAT (Full) requires only 5.4 GB of GPU memory during training, compared to 23.3 GB for full fine-tuning. This 76.8% reduction is not merely a convenience; it is a capability boundary. Full fine-tuning of the 355M parameter RoBERTa-large backbone requires a high-end GPU with 24 GB or more of memory, while LAT can be trained on an entry-level GPU with 8 GB, which is accessible to the vast majority of research groups and commercial teams. The practical implications for cold-start recommendation deployments are substantial: LLM-based recommenders with LAT adaptation can be trained and updated much more frequently and at much lower cost than those requiring full fine-tuning, which directly addresses the challenge of keeping up with rapidly evolving user-item distributions in real recommendation platforms.

The inference throughput of 27.8 tok/ms achieved by LAT is the highest among all compared methods, including full fine-tuning. This is counterintuitive at first glance, since LAT has more total parameters than LoRA and the standard adapter. The explanation lies in the memory efficiency of the LAT configuration: by reducing the training memory requirement, LAT also reduces the working memory occupied by the model during inference, which improves the cache utilisation of the hardware accelerator and reduces the frequency of expensive memory transfer operations. In production recommendation systems where many requests are served concurrently and inference latency is a critical quality-of-service metric, this throughput advantage translates directly into higher service capacity and lower infrastructure cost.

5.6 Synthesis and Theoretical Interpretation

The totality of the experimental results reported in this chapter supports a coherent theoretical account of the determinants of PEFT performance in cold-start recommendation. This account can be organised around a single unifying principle, which this thesis terms the adaptation locality hypothesis: the performance of a parameter-efficient adaptation strategy is determined primarily by the alignment between the layer scope of adaptation and the hierarchical distribution of task-relevant representational modifications in the pretrained model, rather than by the total number of adapted parameters.

This principle has several concrete implications. First, adaptation methods that target the upper layers of the transformer stack, where task-specific semantic representations are concentrated, achieve better performance than methods that adapt uniformly across all layers, because they concentrate their limited adaptation capacity where it is

most needed and avoid the representational damage that lower-layer adaptation causes. Second, the number of adapted parameters is a poor proxy for adaptation quality: LAT with 2.37% of parameters outperforms full fine-tuning with 100% of parameters, while LAT-down with 2.37% of parameters underperforms the standard adapter with 0.45% of parameters. Third, the expressiveness of the adaptation module within each layer matters, but is secondary to the layer scope: removing the nonlinearity from LAT causes smaller performance drops than relocating the adaptation to the lower layers.

This theoretical account also explains the observed failure mode of prompt tuning in demanding recommendation settings. By conditioning the model's computation through the input space without modifying any of its internal weights, prompt tuning effectively asks the pretrained model to solve the task using only its existing representations, redirected through the soft prompt. For tasks where the pretrained representations are already well-aligned with the task requirements, this is sufficient. For tasks like cold-start recommendation, where significant representational realignment is needed to map general semantic knowledge onto domain-specific preference patterns, input-space conditioning is insufficient. The model's internal representations, particularly in the upper layers, must be directly modified, and this requires either weight modification (LoRA or adapters) or the insertion of expressive trainable modules (adapters) at the relevant layer positions.

The cold-start robustness findings add a further dimension to this account by distinguishing the user cold-start and item cold-start failure modes. The item cold-start regime is ultimately a knowledge availability problem: the model's internal representations, however well-adapted they may be, cannot characterise items that are not represented in its training data. Retrieval augmentation addresses this limitation by supplying externally sourced item knowledge at inference time, effectively augmenting the model's parametric knowledge with non-parametric retrieval. The user cold-start regime is a representation alignment problem whose severity scales with the available preference signal: as more interactions become available, the model's ability to personalise improves, and the advantage of LLM-based methods over collaborative filtering narrows but does not disappear. This dual characterisation — item cold-start as a knowledge problem, user cold-start as an alignment problem — provides a principled framework for selecting adaptation strategies in different deployment contexts.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis has presented the first comprehensive, controlled, and unified empirical analysis of parameter-efficient fine-tuning methods for cold-start recommendation, directly addressing a methodological gap in the existing literature that has persisted as a barrier to reliable comparative conclusions about adaptation strategies in this practically important setting. The experimental framework introduced in this thesis — which holds the backbone model, training procedure, and evaluation protocol constant across all compared adaptation methods — provides the clean experimental foundation necessary to attribute observed performance differences to the adaptation mechanism rather than to confounding variables.

The central empirical finding of the thesis is the principle of adaptation locality: performance in cold-start recommendation is determined primarily by where in the model parameters are updated, not by how many parameters are updated. Late Adapter Tuning, which confines adapter modules to the upper layers of the transformer stack and supplements the standard bottleneck architecture with layer normalisation, achieves performance uniformly above full fine-tuning on all four NLP benchmark tasks while updating only 2.37% of total model parameters, reducing GPU memory consumption by 76.8%, and achieving the highest inference throughput of any compared method. This finding contradicts the common implicit assumption that parameter efficiency must trade off against performance quality, and establishes a new design principle for PEFT methods in recommendation: prioritise the identification of the most task-relevant parameter subspaces in the upper model layers, and apply adaptation capacity there with maximum expressiveness and precision, rather than distributing adaptation capacity uniformly across all parameters or all layers.

The secondary findings of the thesis provide important contextual structure around this central result. Prompt-based tuning methods, with fewer than 0.01% of total parameters, achieve the lowest performance on tasks requiring significant representational adaptation, confirming that input-space conditioning is insufficient for highly

personalised and long-tail cold-start recommendation settings. LoRA delivers a strong and reliable efficiency-performance trade-off with 0.22% of parameters, making it the method of choice for settings where LAT's higher parameter count is a constraint. The standard adapter configuration, applied uniformly across all layers, underperforms LAT despite having a similar architectural design, confirming that layer scope rather than adapter architecture is the critical determinant of performance. The cold-start robustness analysis reveals a qualitative asymmetry between user- and item-cold-start failure modes that has direct implications for adaptation strategy selection: item cold-start is best addressed by retrieval-augmented methods that supply external item knowledge, while user cold-start is best addressed by targeted upper-layer adaptation that efficiently aligns pretrained semantic representations with task-specific preference patterns. The ablation study confirms that both the nonlinear expressiveness and the layer normalisation of the LAT adapter modules are necessary components of its performance advantage, and that relocating adaptation to the lower layers rather than the upper layers dramatically degrades performance even with an identical parameter count.

Taken together, these findings reframe parameter-efficient fine-tuning for cold-start recommendation from a compression problem — achieving the same performance as full fine-tuning with fewer parameters — into a representation alignment problem: identifying the minimal and most task-relevant set of parameters that must be updated to align the pretrained model's representations with the task-specific cold-start recommendation objective, and updating those parameters with maximum expressiveness. This reframing has concrete implications for both the design of future PEFT methods and the selection of adaptation strategies for practical deployment.

6.2 Future Work

The findings of this thesis open several productive directions for future research that are grounded in the identified limitations of the current experimental framework and in the theoretical insights developed through the empirical analysis. These directions are discussed below in order of their directness as extensions of the present work.

The most immediately motivated extension is the systematic investigation of hybrid PEFT architectures that combine low-rank updates with adapter modules. The present study's observation that hybrid configurations outperform either component in isolation suggests that LoRA and adapter-based methods capture genuinely complementary aspects of the adaptation space: LoRA captures global, low-rank adjustment of the attention mechanism's directional preferences, while adapters capture lo-

cal, layer-specific nonlinear transformations of the feed-forward representations. A principled hybrid design, in which LoRA is applied to the attention projections of all layers while LAT adapters are applied to the feed-forward sublayers of the upper layers, could potentially achieve better performance than either pure LoRA or pure LAT while maintaining a parameter count well below full fine-tuning. The design space for such hybrid configurations is rich, and exploring it systematically is a natural next step from the present work [8].

The integration of retrieval augmentation with PEFT-based adaptation, motivated by the strong item cold-start results of ColdRAG, represents a second high-value research direction. The present thesis evaluates retrieval augmentation and PEFT adaptation as separate strategies, but their combination — in which retrieval supplies item knowledge that PEFT-adapted upper-layer representations are specifically trained to integrate — has not been investigated. A retrieval-augmented LAT approach, where the model's upper-layer adapters are trained on examples that include externally retrieved item attributes alongside the standard interaction signal, could combine the knowledge coverage advantage of retrieval with the representation alignment advantage of targeted upper-layer adaptation. This combination could be particularly powerful for item cold-start in sparse, long-tail domains like the Office product category studied in this thesis [5].

The scaling behaviour of the adaptation locality principle across different model sizes is a third important research direction. The present experiments use a mid-sized backbone model, and the relative advantage of upper-layer adaptation may be larger or smaller for models at different scales. The hierarchical specialisation of transformer representations is known to be more pronounced in deeper and wider models, which suggests that the adaptation locality principle may become even more important as model scale increases. Evaluating LAT and hybrid PEFT configurations with backbone models of varying scales, from small models suitable for edge deployment to large frontier models, would provide crucial guidance for practitioners working across the full range of computational budgets.

Online and continual adaptation in dynamic recommendation settings represents a fourth research direction that follows directly from the practical motivation of this thesis. Real recommendation platforms must contend with continuously evolving user preferences and a constant stream of new items, meaning that any adaptation strategy must support efficient incremental updating of the model's parameters as new interaction data arrives. PEFT methods, and adapter-based methods in particular, are well-suited to continual adaptation because the frozen backbone provides a stable foundation while the adapter weights can be updated incrementally without risk

of catastrophic forgetting of previously learned representations [13]. Developing and evaluating incremental adaptation protocols for the online cold-start recommendation setting, with explicit guarantees on computational cost per update and on the preservation of previously learned adaptation, is an important step toward practical deployment of LLM-based recommenders in dynamic environments.

Layer-adaptive variants of the LAT approach, in which the set of layers targeted for adaptation is determined dynamically rather than fixed at design time, represent a fifth research direction motivated by the observation that the optimal adaptation scope may vary across tasks, datasets, and individual queries. A learned routing mechanism that uses the input query characteristics to determine which layers require adaptation and applies adapter modules selectively could achieve the efficiency benefits of LAT while adapting the scope of adaptation to the difficulty and domain characteristics of each individual recommendation request. Such adaptive routing has precedents in the mixture-of-experts literature for feed-forward layers, and extending this idea to adapter module placement is a natural and productive extension [12].

Finally, the multi-modal extension of the experimental framework to settings where item representations include visual features, acoustic features, or structured knowledge graph embeddings alongside textual descriptions would substantially broaden the scope and practical relevance of the comparative conclusions. Cold-start recommendation in multi-modal settings, where item metadata is richer but the alignment between different modalities and user preference patterns is more complex, may favour different PEFT strategies than the text-only settings studied in this thesis. The development of multi-modal PEFT methods that can leverage all available item feature types efficiently and effectively, evaluated under the controlled benchmark framework introduced in this thesis, provides a coherent and ambitious research programme for the next generation of cold-start LLM-based recommender systems.

BIBLIOGRAPHY

[1] B. Hao, H. Yin, J. Zhang, C. Li, and H. Chen, "A multi-strategy-based pre-training method for cold-start recommendation," *ACM Transactions on Information Systems*, vol. 41, no. 2, pp. 31:1–31:24, Jan. 2023.

[2] T. Li *et al.*, "Memory-augmented meta-learning on meta-path for fast adaptation cold-start recommendation," *Connection Science*, vol. 34, no. 1, pp. 301–318, Dec. 2022.

[3] H. Lyu *et al.*, "LLM-rec: Personalized recommendation via prompting large language models," in *Findings of the ACL: NAACL 2024*, 2024, pp. 583–612.

[4] R. Kumar, T. Dey, and S. Das, "Empowering recommendations: Survey of LLM-based recommendation systems," in *2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, Jan. 2025, pp. 1–6.

[5] Y. Jiang *et al.*, "Prompt tuning for item cold-start recommendation," arXiv:2412.18082, Dec. 2024.

[6] ——, "Prompt tuning for item cold-start recommendation," in *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24)*. ACM, Oct. 2024, pp. 411–421.

[7] T. Li *et al.*, "Memory-augmented meta-learning on meta-path for fast adaptation cold-start recommendation," *Connection Science*, vol. 34, no. 1, pp. 301–318, Dec. 2022.

[8] Z. Yang, H. Lin, J. Xue, and Z. Zhang, "GR-LLMs: Recent advances in generative recommendation based on large language models," arXiv:2507.06507, Jul. 2025.

[9] K. Büyükakyüz, "OLoRA: Orthonormal low-rank adaptation of large language models," arXiv:2406.01775, Nov. 2024.

[10] Sudha, A. Sagu, and N. P. Singh, "LLM-empowered recommender systems: Next-generation frameworks and applications," in *2025 2nd International Conference on Advanced Computing and Emerging Technologies (ACET)*, Nov. 2025, pp. 1–4.

[11] Z. Yi, I. Ounis, and C. MacDonald, "Contrastive graph prompt-tuning for cross-domain recommendation," *ACM Transactions on Information Systems*, vol. 42, no. 2, pp. 60:1–60:28, Dec. 2023.

[12] Y. Zhao *et al.*, "Meta-learning for cold-start personalization in prompt-tuned LLMs," arXiv:2507.16672, Jul. 2025.

[13] Y. Luo *et al.*, "Online item cold-start recommendation with popularity-aware meta-learning," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Jul. 2025, pp. 927–937.

[14] K. Weimann and T. O. F. Conrad, "FELRec: Efficient handling of item cold-start with dynamic representation learning in recommender systems," arXiv:2210.16928, Oct. 2024.

[15] Anonymous, "Few-shot learning for cold-start recommendation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024. [Online]. Available: https://aclanthology.org/2024.lrec-main.631/

[16] X. Wu *et al.*, "Could small language models serve as recommenders? towards data-centric cold-start recommendations," arXiv:2306.17256, Mar. 2024.