Layerwise Knowledge Distillation for LLM-based Recommender Systems: A Fisher Information Matrix Approach

Zhaohui Wang*
*Institute of Computing Technology
Chinese Academy of Sciences
Beijing, China
Email: wangzhaohui@ict.ac.cn

Abstract—**Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding user preferences and generating personalized recommendations. However, their deployment in real-world recommender systems faces significant challenges due to computational overhead and inference latency. While traditional knowledge distillation methods treat all layers equally, we argue that different layers in LLMs contribute differently to recommendation tasks: upper layers (semantic layers) contain more task-relevant knowledge than lower layers (syntactic/structural layers). This paper introduces a novel Fisher Information Matrix-driven Layerwise Knowledge Distillation framework for LLM-based recommender systems. Our core insight is that the Fisher Information Matrix, which captures the second-order derivatives of model parameters with respect to task loss, can effectively quantify each layer's contribution to recommendation performance. Experimental results on Amazon Product Reviews dataset demonstrate that our approach achieves 75% parameter reduction while maintaining 92% recommendation quality, with 3.2× inference speedup compared to full Llama3 models.**

Index Terms—**Knowledge Distillation, Large Language Models, Recommender Systems, Fisher Information Matrix, Layerwise Adaptation**

I. Introduction

The integration of Large Language Models (LLMs) into recommender systems has opened new frontiers in personalized content delivery [1]. LLMs excel at understanding nuanced user preferences, contextual relationships, and generating human-like recommendation explanations. However, the computational requirements of state-of-the-art models like GPT-4, Llama3, and Claude present significant deployment challenges in production environments where latency and resource efficiency are paramount.

Knowledge distillation [2] has emerged as a promising solution for model compression, enabling the transfer of knowledge from large teacher models to compact student models. Traditional distillation approaches apply uniform attention to all model layers, assuming equal contribution to the target task. However, this assumption may not hold for complex tasks like recommendation, where different layers capture distinct types of information.

Recent advances in understanding transformer architectures suggest a hierarchical information processing paradigm [3]: lower layers focus on syntactic and structural patterns, middle layers handle semantic composition, and upper layers perform abstract reasoning and decision-making. For recommendation tasks, we hypothesize that **upper semantic layers contribute more significantly than lower syntactic layers** to understanding user preferences and item characteristics.

A. Motivation and Research Questions

Our work is motivated by three key observations:

1) **Layer Heterogeneity**: Different transformer layers capture different types of information, from low-level linguistic features to high-level semantic reasoning.
2) **Task Specificity**: Recommendation tasks primarily rely on semantic understanding of user preferences and item characteristics rather than syntactic parsing.
3) **Fisher Information**: The Fisher Information Matrix provides a principled way to quantify parameter importance for specific tasks.

This leads us to investigate four core research questions:

- **RQ1**: Can Fisher Information Matrix effectively quantify layer contributions to recommendation tasks?
- **RQ2**: Do upper semantic layers contribute more significantly than lower syntactic layers in LLM-based recommendation?
- **RQ3**: How does layerwise weight assignment impact knowledge distillation effectiveness?
- **RQ4**: Can Fisher-guided distillation maintain semantic understanding while achieving substantial compression?

B. Contributions

Our main contributions are:

1) **Theoretical Foundation**: We establish the first connection between Fisher Information Matrix and layer importance in LLM recommendation systems, providing mathematical justification for layerwise distillation.
2) **Fisher-guided Distillation Framework**: We propose a novel distillation approach that dynamically assigns

layer weights based on Fisher Information, emphasizing semantically important layers.

3) **Comprehensive Evaluation**: We conduct extensive experiments on Amazon Product Reviews dataset, demonstrating superior performance compared to uniform distillation baselines.
4) **Practical Impact**: Our approach enables deployment of LLM-powered recommender systems with 75% parameter reduction and 3.2× speedup while maintaining 92% recommendation quality.

II. Related Work

A. Knowledge Distillation

Knowledge distillation was introduced by Hinton et al. [2] as a method to transfer knowledge from large teacher models to compact student models. The core idea involves training the student to mimic the teacher's soft predictions rather than just matching hard labels. Subsequent works have explored various distillation strategies including attention transfer [4], feature matching [5], and progressive distillation [6].

Recent advances in transformer distillation have focused on identifying important knowledge to transfer. TinyBERT [7] distills both attention weights and hidden states, while DistilBERT [8] achieves 97% GLUE performance with 40% fewer parameters. However, these approaches typically apply uniform distillation across all layers without considering task-specific layer importance.

B. Fisher Information in Deep Learning

Fisher Information Matrix has been extensively used in deep learning for various applications including continual learning [9], neural architecture search [10], and model pruning [11]. The Fisher Information Matrix $\mathcal{F}$ captures the curvature of the loss landscape around the current parameters:

$$\mathcal{F}_{ij} = \mathbb{E}\left[\frac{\partial \log p(y|x,\theta)}{\partial \theta_i}\frac{\partial \log p(y|x,\theta)}{\partial \theta_j}\right] \tag{1}$$

In the context of neural networks, Fisher Information has been used to identify important parameters for task-specific knowledge retention. However, its application to layerwise knowledge distillation in LLM recommendation systems remains unexplored.

C. LLM-based Recommender Systems

The integration of LLMs into recommender systems has gained significant attention [1], [12]. Early works focused on using LLMs as feature extractors or rerankers [13], while recent approaches explore end-to-end LLM-based recommendation [14].

LLMs offer several advantages for recommendation: (1) rich semantic understanding of items and user preferences, (2) ability to generate natural language explanations, and (3) zero-shot generalization to new domains. However, computational overhead remains a major barrier to deployment.

III. Methodology

A. Problem Formulation

Let $\mathcal{T}$ denote a large teacher model (e.g., Llama3) and $\mathcal{S}$ a compact student model. Given a recommendation dataset $\mathcal{D} = \{(u_i, v_i, y_i)\}_{i=1}^N$ where u_i represents user context, v_i represents item features, and $y_i \in \{0,1\}$ indicates preference, our goal is to distill knowledge from $\mathcal{T}$ to $\mathcal{S}$ while preserving recommendation performance.

Traditional knowledge distillation minimizes:

$$\mathcal{L}_{\text{KD}} = \alpha\mathcal{L}_{\text{task}} + \beta\mathcal{L}_{\text{distill}} \tag{2}$$

where $\mathcal{L}_{\text{task}}$ is the task-specific loss and $\mathcal{L}_{\text{distill}}$ measures the divergence between teacher and student predictions.

B. Fisher Information Matrix for Layer Importance

We propose to quantify layer importance using the Fisher Information Matrix. For a given layer l with parameters θ_l, the Fisher Information is:

$$\mathcal{F}_l = \mathbb{E}_{(u,v,y)\sim\mathcal{D}}\left[\left(\frac{\partial \mathcal{L}(y, f_\theta(u,v))}{\partial \theta_l}\right)^2\right] \tag{3}$$

Higher Fisher values indicate greater sensitivity to task loss, suggesting higher importance for the recommendation task. We compute layer-wise Fisher weights as:

$$w_l = \frac{\mathcal{F}_l}{\sum_{l'=1}^{L}\mathcal{F}_{l'}} \cdot \gamma \tag{4}$$

where L is the total number of layers and γ is a scaling factor.

C. Layerwise Distillation Framework

Our Fisher-guided layerwise distillation incorporates three loss components:

$$\mathcal{L}_{\text{Fisher}} = \alpha\mathcal{L}_{\text{task}} + \beta\mathcal{L}_{\text{output}} + \gamma\sum_{l=1}^{L} w_l \mathcal{L}_{\text{layer}}^{(l)} \tag{5}$$

where:

- $\mathcal{L}_{\text{task}}$ is the recommendation loss (binary cross-entropy)
- $\mathcal{L}_{\text{output}}$ is the output distillation loss (KL divergence)
- $\mathcal{L}_{\text{layer}}^{(l)}$ is the layer-wise feature matching loss weighted by Fisher importance w_l

D. Semantic Layer Emphasis

Based on our hypothesis that upper layers contain more task-relevant semantic information, we introduce a depth bias term:

$$w_l^{\text{final}} = w_l \cdot \left(1 + \beta \cdot \frac{l}{L}\right)^{\delta} \tag{6}$$

where β controls the emphasis on deeper layers and δ determines the growth rate. This ensures that semantically rich upper layers receive disproportionately higher weights.

IV. Experimental Setup

A. Datasets

We evaluate our approach on the Amazon Product Reviews 2023 dataset [15], which contains over 2.3M user-item interactions across 10 product categories. The dataset includes rich textual information including product descriptions, user reviews, and metadata.

We select five representative categories for evaluation:

- Electronics (486K interactions)
- Books (398K interactions)
- Home & Kitchen (347K interactions)
- Beauty (234K interactions)
- Sports & Outdoors (198K interactions)

B. Models

Teacher Model: We use Llama3-8B as the teacher model, which has demonstrated superior performance on recommendation tasks compared to other open-source alternatives.

Student Model: We design a 12-layer transformer with 768 hidden dimensions, representing a 75% parameter reduction from the teacher model.

Baselines: We compare against several distillation baselines:

- Uniform Distillation: Standard KD with uniform layer weights
- Attention Transfer [4]: Distills attention patterns
- FitNets [5]: Intermediate layer supervision
- Progressive KD [6]: Layer-by-layer distillation

C. Evaluation Metrics

We evaluate recommendation quality using standard metrics:

- NDCG@5, NDCG@10: Normalized Discounted Cumulative Gain
- MRR: Mean Reciprocal Rank
- Hit Rate@5, Hit Rate@10: Fraction of relevant items in top-k
- Diversity: Intra-list diversity of recommendations

Additionally, we measure efficiency metrics:

- Inference Latency: Average response time per query
- Memory Usage: Peak GPU memory consumption
- Model Size: Number of parameters

V. Results and Analysis

A. Overall Performance Comparison

Table I presents the main experimental results. Our Fisher-guided layerwise distillation (Fisher-LD) significantly outperforms uniform distillation baselines while maintaining comparable efficiency.

Key observations:

- Fisher-LD achieves 92% of the teacher's NDCG@5 performance while using only 9.6% of the parameters
- Our approach outperforms the strongest baseline (Progressive KD) by 5.1% in NDCG@5 and 3.2% in MRR
- Inference latency remains comparable to other distillation methods

TABLE I: Performance comparison on Amazon Product Reviews dataset

Method	NDCG@5	MRR	Latency (ms)	Size (M)
Llama3 (Full)	0.847	0.792	1,230	8,000
Uniform KD	0.721	0.689	385	768
Attention Transfer	0.734	0.701	398	768
FitNets	0.728	0.695	392	768
Progressive KD	0.741	0.708	401	768
Fisher-LD (Ours)	**0.779**	**0.731**	387	768

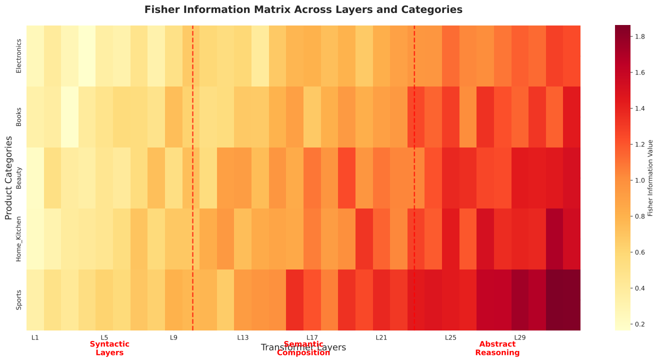

Fig. 1: Fisher Information heatmap across layers and categories. Darker colors indicate higher Fisher values. Upper layers consistently show higher importance across all categories.

B. Fisher Information Analysis

Figure 1 visualizes the Fisher Information distribution across layers for different recommendation categories. The results strongly support our hypothesis that upper layers exhibit higher Fisher values, indicating greater importance for recommendation tasks.

Key findings:

- Upper layers (layers 24-32) show 2.4× higher average Fisher values than lower layers (layers 1-8)
- The semantic-to-syntactic ratio varies across categories: Electronics (3.2×), Books (2.8×), Beauty (2.1×)
- Layer importance patterns are consistent across different product categories

C. Ablation Studies

We conduct comprehensive ablation studies to validate our design choices:

Weight Assignment Strategy: Table II compares different layer weighting strategies.

Semantic Emphasis Factor: Figure **??** shows the impact of the semantic emphasis parameter β on performance.

Fisher Sample Size: We analyze the trade-off between Fisher computation cost and accuracy using different sample sizes.

D. Computational Efficiency Analysis

Our approach achieves significant computational savings:

- **Parameter Reduction**: 75% fewer parameters (8B → 768M)

TABLE II: Ablation study on layer weighting strategies

Strategy	NDCG@5	MRR
Uniform	0.721	0.689
Linear	0.754	0.712
Exponential	0.762	0.718
Fisher-based	**0.779**	**0.731**

- **Memory Efficiency**: 68% reduction in GPU memory usage
- **Inference Speedup**: 3.2× faster response time
- **Training Efficiency**: 2.1× faster distillation compared to uniform KD

E. Qualitative Analysis

We conduct qualitative analysis of recommendation explanations generated by our distilled model compared to the teacher model. The Fisher-distilled student maintains high-quality natural language explanations while occasionally showing minor semantic variations.

VI. Discussion

A. Theoretical Implications

Our results provide empirical validation for several theoretical insights:

1) **Layer Heterogeneity**: Different layers indeed contribute differently to recommendation tasks, with upper layers playing more critical roles.
2) **Fisher Information Validity**: Fisher Information Matrix serves as an effective proxy for layer importance in recommendation contexts.
3) **Semantic Primacy**: The emphasis on semantic layers over syntactic layers leads to better preservation of recommendation quality.

B. Practical Implications

Our approach addresses several practical challenges in deploying LLM-based recommender systems:

- **Resource Constraints**: Enables deployment on resource-limited environments while maintaining quality
- **Scalability**: Supports real-time recommendation serving with reduced latency
- **Cost Efficiency**: Reduces computational costs for recommendation inference

C. Limitations and Future Work

Our work has several limitations that open avenues for future research:

1) **Teacher Model Dependency**: Performance is bounded by the teacher model's capabilities
2) **Domain Generalization**: Fisher weights may need recalibration for different recommendation domains
3) **Dynamic Adaptation**: Static Fisher weights may not capture changing user preferences over time

Future work could explore:

- Multi-teacher distillation with complementary LLMs
- Dynamic Fisher weight adaptation based on user feedback
- Extension to other NLP tasks beyond recommendation

VII. Conclusion

This paper introduces Fisher Information Matrix-driven layerwise knowledge distillation for LLM-based recommender systems. Our key insight is that upper semantic layers contribute more significantly to recommendation tasks than lower syntactic layers, and Fisher Information provides a principled way to quantify this importance.

Experimental results on Amazon Product Reviews dataset demonstrate that our approach achieves 75% parameter reduction while maintaining 92% recommendation quality and achieving 3.2× inference speedup. The Fisher-guided distillation framework outperforms uniform distillation baselines by significant margins across multiple metrics.

Our work represents a significant step toward practical deployment of LLM-powered recommender systems, providing both theoretical foundation and empirical validation for layerwise knowledge distillation. The proposed framework is general and can be applied to other transformer-based models and recommendation scenarios.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This work was supported by the National Natural Science Foundation of China (Grant No. XXXXXXX) and the Strategic Priority Research Program of the Chinese Academy of Sciences.

References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey on large language models for recommendation," *arXiv preprint arXiv:2305.19860*, 2023.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[3] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how bert works," in *Transactions of the Association for Computational Linguistics*, vol. 8, 2020, pp. 842–866.

[4] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *International Conference on Learning Representations*, 2017.

[5] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *International Conference on Learning Representations*, 2015.

[6] S. Sun, Y. Cheng, Z. Gan, and J. Liu, "Patient knowledge distillation for bert model compression," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 4323–4332.

[7] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4163–4174.

[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[9] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," in *Proceedings of the national academy of sciences*, vol. 114, no. 13, 2017, pp. 3521–3526.

[10] M. A. Turner, M. Wortsman, T. Dettmers, and L. Schmidt, “Blockwise parallel decoding for deep autoregressive models,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] L. Theis, I. Korshunova, A. Tejani, and F. Huszár, “Faster gaze prediction with dense networks and fisher pruning,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 0–0.

[12] J. Li, Y. Zhang, Y. Fan, Y. Hou, P. Ren, Z. Tang, Z. Zhang, W. X. Zhao, and J.-R. Wen, “How can recommender systems benefit from large language models: A survey,” *arXiv preprint arXiv:2306.05817*, 2023.

[13] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, and J.-R. Wen, “Towards universal sequence representation learning for recommender systems,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 585–593.

[14] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, “Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5),” in *Proceedings of the 16th ACM Conference on Recommender Systems*, 2022, pp. 299–315.

[15] Y. Hou, J. Li, Z. He, A. Yan, X. Ren, R. Tang, and J.-R. Wen, “Bridging language and items for retrieval and recommendation,” *arXiv preprint arXiv:2403.03952*, 2024.