Align Before You Recommend: Parameter Efficient Personalization via Cross-Attentive Fusion of Hierarchical Language Models

Alicja Kwasniewska
Amazon Web Services
akws@amazon.com

Marcin Bednarz
Amazon Web Services
mbz@amazon.com

Chad Neal
Amazon Web Services
chadneal@amazon.com

Abstract

The rapidly growing global advertising and marketing industry demands innovative machine learning systems that balance accuracy with efficiency. Recommendation systems, crucial to many platforms, require careful considerations and potential enhancements. While Large Language Models (LLMs) have transformed various domains, their potential in sequential recommendation systems remains underexplored. Pioneering works like Hierarchical Large Language Models (HLLM) demonstrated LLMs' capability for next-item recommendation but rely on computationally intensive fine-tuning, limiting widespread adoption.

This work introduces HLLM+, enhancing the HLLM framework to achieve high-accuracy recommendations without full model fine-tuning. By introducing targeted alignment components between frozen LLMs, our approach outperforms frozen model performance in popular and long-tail item recommendation tasks by 29% while reducing training time by 29%. We also propose a ranking-aware loss adjustment, improving convergence and recommendation quality for popular items. Experiments show HLLM+ achieves superior performance with frozen item representations allowing for swapping embeddings, also for the ones that use multimodality, without tuning the full LLM. These findings are significant for the advertising technology sector, where rapid adaptation and efficient deployment across brands are essential for maintaining competitive advantage.

1. Introduction

In the rapidly expanding global advertising market, projected to surge from $676.8 billion in 2024 to $995 billion by 2033 [8], there is an urgent need to enhance the algorithmic foundations that power these systems. Sequential recommendation algorithms stand at the forefront of this evolution, serving as the backbone for a wide range of critical applications in digital advertising and marketing. These algorithms drive personalized content delivery [17], enabling platforms to tailor user experiences in real-time. They power sophisticated targeted marketing campaigns [7], allowing advertisers to reach the right audience with precision. Moreover, these algorithms play a crucial role in mapping and optimizing the consumer journey [23], helping businesses understand and influence customer behavior across digital platforms. As digital advertising grows more complex, better recommendation algorithms are needed to maintain competitive edge and maximize return on investment in digital marketing strategies.

Sequential recommendation is a well-studied task that aims to model each user as a sequence of items interacted in the past and predict top-K ranked items that a user will likely interact with in a near future. Various techniques have already been proposed in the literature to tackle sequential recommendation, including traditional machine learning approaches, such as matrix factorization with Markov chains [21], as well as deep learning methods, e.g., recurrent neural networks (e.g., GRU4Rec [6]), transformer-based models with self-attention [10], bidirectional encoder representations (e.g., BERT4Rec [24]), or graph-based session recommenders (e.g., SR-GNN [28]).

In recent years, Large Language Models (LLMs) have transformed various domains through their extensive language understanding and generation capabilities. However, their application in recommendation systems, particularly sequential recommendation, remains largely unexplored. While transformer and attention-based models have been successfully applied to recommender systems [10, 31], the application of Large Language Models (LLMs) to this domain is relatively recent.

Initial studies approached the problem as a conversational task, employing instruction-following and prompt-engineering techniques for next-item prediction [15, 27]. Subsequent enhancements incorporated additional metadata, such as ID-based item embeddings [13]. A common approach has been to use pre-trained LLMs for encoding item representations [9], user profiles [14], or collaborative knowledge [20].

A key research question is whether the knowledge accumulated during LLM pretraining can be effectively leveraged for modeling item-user relationships. The Hierarchical Large Language Model (HLLM) [3], pioneered by Li et al., investigated this question by examining whether off-the-shelf LLMs, pre-trained on massive corpora, could be effectively used for recommendations with or without fine-tuning. HLLM was introduced to address a limitation of many previous studies, which offered only marginal improvements over traditional systems. This was largely due to the incomplete leveraging of LLMs for item/user bridging in end-to-end frameworks.

While these models demonstrate strong general knowledge, the necessity and effectiveness of task-specific fine-tuning for recommendation systems remained an open question. The performed experiments concluded that fine-tuning is necessary for optimal performance in recommendation tasks. HLLM demonstrated that fully fine-tuned LLMs outperform the frozen setting significantly, suggesting that pre-trained language representations, while powerful, require task-specific adaptation for recommendation.

However, full fine-tuning of large language models presents substantial computational challenges, requiring extensive resources and time. In this work, we challenge the necessity of comprehensive fine-tuning by proposing HLLM+, an enhanced approach that leverages representation alignment techniques to bridge the gap between language understanding and recommendation tasks while keeping the core LLM parameters frozen.

Our key insight is that the fundamental issue lies not in the quality of LLM representations but in the alignment between different representation spaces for user and item embeddings. We propose that through carefully designed cross-attention mechanisms and projection layers, we can effectively align item and user representations without modifying the underlying language models. This approach preserves the rich semantic knowledge captured during pre-training while enabling effective recommendation-specific adaptation. It not only reduces computational requirements but also potentially mitigates catastrophic forgetting, where fine-tuning on recommendation tasks might compromise the general language understanding capabilities of LLMs.

Specifically, this work proposes the following innovations:

1. HLLM+ model enhancement that aligns item and user representations, eliminating the need for full LLM fine-tuning
2. A novel alignment approach that enables integration of multimodal item LLMs with differently-architected user LLMs, bridging the gap between multimodal item descriptions and user preference modeling
3. A novel loss adjustment focused on ranking performance, enabling better convergence with frozen representations
4. An investigation demonstrating the framework's compatibility with smaller LLMs without architectural modifications

Our experiments show that the proposed HLLM+ enhancement allows for outperforming frozen model setting on popular and long-tail item recommendation tasks. The rest of the work is structured as follows: Section 2 presents related work for LLMs in recommendation tasks, Section 3 introduces methodology proposed to enhance HLLM model with the alignment objective. Experiments and results are shown in Section 4 along with Ablation Study. The work is discussed in Section 5 and concluded in Section 6.

2. Method

This section presents a novel framework that leverages two Large Language Models (LLMs) interconnected through a specialized alignment layer. This layer facilitates efficient integration of language comprehension capabilities with recommendation-specific tasks. We begin with the problem formulation, followed by an explanation of how the Hierarchical Large Language Model (HLLM) [3] approach adapts LLMs to recommendation tasks. Finally, we introduce the HLLM+ enhancement, which enables efficient alignment between item and user representations without requiring extensive fine-tuning of the underlying language models.

2.1. Problem Formulation

The sequential recommendation task involves predicting a user's next interaction based on their historical sequence of interactions. Given a user $u \in \mathcal{U}$ and their chronological interaction sequence $\mathcal{S}u = I_1, I_2, ..., I_n$ where each I_i represents an item from the item set $\mathcal{I}$, the goal is to predict the next item I_{n+1} that the user is likely to interact with. Each item I has its corresponding ID and text information (e.g., title, tags, etc.), and the proposed HLLM+, same as HLLM, leverages this text information through language models.

2.1.1. HLLM for Recommendation

HLLM leverages two large language models: an item LLM that captures rich item semantics and a user LLM that models sequential user preferences. The original HLLM approach demonstrated that while LLMs provide powerful semantic understanding, full fine-tuning of both models is necessary to align them with recommendation objectives.

2.1.2. HLLM+ Enhanced Model

Extensive fine-tuning of language models is resource-intensive, requiring significant time, computational power, and financial investment. Furthermore, many organizations lack access to specialized training hardware, and in the context of item recommendation, the rapid introduction of new items makes continuous full model retraining impractical.

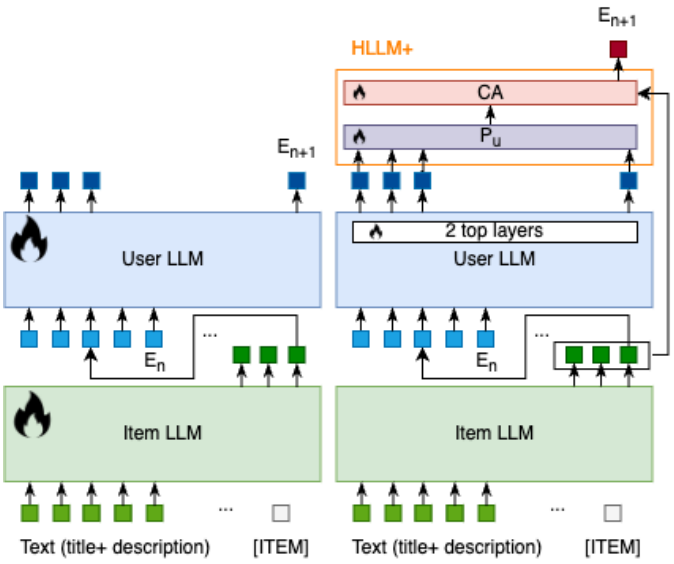

Figure 1. Comparison of HLLM architectures. (Left) Original HLLM with trainable Item and User LLMs, where Item LLM converts text descriptions into embeddings and appends [ITEM] token. (Right) Enhanced HLLM+ with frozen LLMs and additional projection and cross-attention components (highlighted in orange), enabling efficient adaptation without full model fine-tuning.

HLLM+ offers a solution to these challenges by maintaining frozen core language models while introducing targeted alignment components. These components efficiently connect language understanding with recommendation tasks, providing a more adaptable and cost-effective approach. As shown in Fig. 1, we maintain HLLM hierarchical structure where the Item LLM first converts item descriptions into embeddings using the same prompting strategy, followed by the User LLM which processes these embeddings to model user interests. Our enhancement introduces additional layers between these two components to better align the representations, while keeping the core functionality and input/output format of both LLMs unchanged.

Let $\mathbf{E}_u = e_1, e_2, \ldots, e_n$ represent the embeddings of items in the user sequence, where $e_i \in \mathbb{R}^d$ is the embedding of item I_i derived from an item language model. HLLM+ enhanced sequential recommendation task is defined as:

$$\mathbf{H}_u = \text{Enc}_U(\mathbf{E}_u) \tag{1}$$

$$\mathbf{P}_u = \text{Proj}(\mathbf{H}_u) \tag{2}$$

$$\mathbf{Z}_u = \text{CA}(\mathbf{P}_u, \mathbf{E}_u, \mathbf{M}_u) \tag{3}$$

$$\hat{I}_{n+1} = \arg\max I \in \mathcal{I}\tau \cdot \cos(\mathbf{Z}_u, \mathbf{E}_I) \tag{4}$$

where Enc_U is the frozen user encoder language model, Proj is an enhanced projection layer with residual connections, CA is a cross-attention mechanism, $\mathbf{M}_u$ is an attention mask for the user sequence, and τ is the temperature parameter. The projection layer Proj transforms the LLM's hidden representations into a recommendation-friendly space while preserving semantic knowledge. This addresses the representation gap between language modeling and recommendation objectives. The cross-attention mechanism CA enables dynamic focus on relevant parts of interaction history, creating a more nuanced user representation that captures both sequential patterns and item relationships. Cross-attention is particularly powerful as it allows the model to learn which historical interactions are most predictive of future preferences without modifying the underlying LLM parameters. Together, these lightweight components enable effective alignment of frozen LLM representations for recommendation tasks with significantly reduced computational requirements.

2.2. Enhancement Details

This section outlines the proposed key enhancements that restore recommendation accuracy for popular items via representation alignment while avoiding the burden of full model fine-tuning. The projection layer in Equations 2 consists of two sequential projections with residual connections and layer normalization:

$$\mathbf{P}_{u,1} = \mathbf{H}_u + \text{Dropout}(\phi(\mathbf{W}_1 \cdot \text{Norm}(\mathbf{H}_u))) \tag{5}$$

$$\mathbf{P}u = \mathbf{P}u, 1 + \text{Dropout}(\phi(\mathbf{W}2 \cdot \text{Norm}(\mathbf{P}u, 1))) \tag{6}$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d \times d}$ are learnable parameter matrices, Norm is layer normalization, ϕ is the GELU activation function, and Dropout is the dropout regularization.

2.2.1. Cross-attention Mechanism

HLLM+ cross-attention mechanism is designed for computational efficiency while effectively aligning user and item representations:

$$\mathbf{Q} = \mathbf{W}^Q\mathbf{P}_u, \quad \mathbf{K} = \mathbf{W}^K\mathbf{E}_u, \quad \mathbf{V} = \mathbf{W}^V\mathbf{E}_u \tag{7}$$

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \tag{8}$$

$$\mathbf{C} = \mathbf{A}\mathbf{V} \tag{9}$$

$$\mathbf{Z}_u = \mathbf{P}_u + \sigma(g) \cdot \text{Dropout}(\mathbf{C}) \tag{10}$$

$$\mathbf{Z}_u = \text{Norm}(\mathbf{Z}_u) \tag{11}$$

where $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d \times d}$ are learnable projection matrices, g is a learnable gate parameter, and $\sigma(\cdot)$ is the sigmoid function. This simple cross-attention approach allows the model to selectively attend to relevant parts of the user's interaction history based on the current user representation. The gating mechanism enables the model to gradually learn how much cross-attention information to incorporate, starting with a conservative bias toward the original representation. This design is particularly effective when working with frozen language models, as it provides a controlled way to align representations without disrupting the pre-trained knowledge.

2.3. Training Objective

HLLM+ is trained with a combined loss function that incorporates both noise contrastive estimation (NCE)[18] and a custom ranking-aware components:

$$\mathcal{L} = \mathcal{L}_{\text{NCE}} + \lambda_r \mathcal{L}_{\text{rank}} \tag{12}$$

The NCE component is defined as:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{|\mathcal{B}|} \sum (u, i^+) \in \mathcal{B} \log \frac{e^{\tau \cdot s_{u,i^+}}}{e^{\tau \cdot s_{u,i^+}} + \sum_{i^- \in \mathcal{N}} e^{\tau \cdot s_{u,i^-}}} \tag{13}$$

where $s_{u,i} = \cos(\mathbf{Z}_u, \mathbf{E}_i)$ is the cosine similarity between user representation and item embedding, τ is the temperature parameter, $\mathcal{B}$ is the batch of user-positive item pairs, and $\mathcal{N}$ is the set of negative items.

The ranking-aware loss provides an additional regularization effect by focusing on hard negative examples:

$$\mathcal{L}_{\text{rank}} = \frac{1}{|\mathcal{B}|} \sum (u, i^+) \in \mathcal{B} \frac{1}{k} \sum_{j=1}^{k} \sigma_+(s(u, i_j^-) - s(u, i^+)) \tag{14}$$

where:

$$\sigma_+(x) = \log(1 + e^x) \tag{15}$$

is a softplus function used instead of a traditional hinge loss to ensure smooth gradient and continuous penalization of the model based on how close the negative items are to the positive items in the embedding space. i_j^- are the top-k hard negative items with highest similarity to the user representation. The weight λ_r increases gradually during training through a warm-up schedule. This approach offers two key benefits: (1) it helps the model better distinguish between semantically similar but incorrect items and correct recommendations, and (2) it serves as an effective regularization technique, leading to better generalization across popular and long-tail items. By gradually increasing λ_r, we allow the model to first learn general patterns through NCE before refining its ability to make more precise distinctions.

For inference, HLLM+ uses a mean pooling strategy that balances recency bias with historical context:

$$\mathbf{Z}u^{\text{pool}} = \frac{\sum_{j=1}^{n} m_j \mathbf{Z}_{u,j}}{\sum_{j=1}^{n} m_j} \tag{16}$$

where $m_j \in 0, 1$ indicates whether position j contains a valid item in the user sequence.

While cross-attention already provides a mechanism for the model to focus selectively on different parts of the input sequence, the pooled representation serves a complementary purpose during inference. Cross-attention creates position-specific contextualized representations that account for interactions between items in the sequence. However, these representations still maintain position-specific information.

The pooling operation aggregates information across all positions, creating a more comprehensive user representation that captures broader patterns beyond what any single position might reflect. This is particularly important because cross-attention may emphasize certain positions over others based on learned relevance patterns, potentially deemphasizing older interactions that might still be relevant for long-tail recommendations.

3. Experiments

The effectiveness of the proposed HLLM+ enhancement in aligning off-the-shelf LLMs to the sequence recommendation task is evaluated through a series of experiments using the Movie Lens 1M Ratings Dataset [5]. Due to the extensive nature of experiments, a subset of the dataset (6020 random users, comprising 104327 interactions, and 201 items) is utilized to compare the proposed enhancements against the baseline model. Since the model architecture requires item descriptions, they were added synthetically during preprocessing.

3.1. Quantitative Evaluation

A key advantage of the HLLM model is its architecture-agnostic nature for both user and item models. While previous HLLM experiments demonstrated improved accuracy with larger models, our evaluation focuses on a different aspect. Using the smaller SmolLM model [1], we investigate two key points: 1) whether smaller models retain sufficient pre-trained knowledge for recommendation tasks, and 2) how the alignment layer impacts recommendation accuracy compared to frozen models without such alignment, rather than measuring absolute accuracy performance.

We train the original HLLM model on the Movie Lens using 101M parameter Smol variant (smol_llama101M-

GQA[19]) for both user and item models under two configurations: Both models fully tunable, as recommended in the original work; Item model frozen with user model learnable, which showed lowest performance in HLLM-1B Ablation studies on Pixel200K.

For the second configuration, we limit fine-tuning to the top 2 layers of the user model to maintain consistency across experiments. Additionally, we evaluate HLLM+ by training its alignment layer while maintaining this same configuration. Our evaluation focuses on popular item recommendations, a critical setting in digital advertising and marketing platforms, where accurately recommending frequently purchased or trending items directly impacts business metrics and user engagement. Hyperparameters for model tuning were set to defaults, i.e., learning rate of 1e-4, epochs 5, max_text_length 256, max_item_list_length 10. Number of negatives was set to 128. Performance was measured using standard recommendation metrics: Recall@K (proportion of relevant items in the top-K recommendations), hereafter referred to as R@, and NDCG (Normalized Discounted Cumulative Gain, which measures ranking quality by assigning higher weights to relevant items appearing earlier in recommendations), hereafter referred to as N@. Experiments were performed on a g5.2xlarge instance with 1 24 GB GPU. Results are presented in Table 1 (train batch size 4, MAX TEXT LENGTH 256, MAX ITEM LIST LENGTH 10), training efficiency is shown in Table 2.

Model	Tunability		Metrics			
	Item	User	R@5	R@10	N@5	N@10
HLLM	All	All	37.78	53.84	24.83	30.01
HLLM	Frozen	2 top	17.50	27.52	11.43	16.48
HLLM+	Frozen	2 top	22.73	34.20	14.75	18.36

Table 1. Performance comparison of HLLM and HLLM+ models

Model	Configuration	Time/Epoch (s)	Loss
HLLM	All layers tunable	1505	3.70
HLLM	Item frozen, User 2 top	524	4.87
HLLM+	Item frozen, User 2 top	1066	4.37

Table 2. Training efficiency comparison showing time per epoch and training loss

Based on these results, the HLLM+ model demonstrates improvements over the baseline HLLM configurations. When comparing models with frozen item embeddings and only the top two user layers tunable, HLLM+ outperforms HLLM across all metrics, with notable improvements in R@5 (29% increase), and N@5 (29% increase). However, we also found out that HLLM+ can't match the performance of the fully tunable HLLM model in most metrics on Movie Lens dataset.

In terms of computational efficiency, HLLM+ offers substantial benefits. It reduces training time per epoch by 29% compared to the fully tunable HLLM (1066s vs 1505s). The combination of improved accuracy, reduced training time, and lower loss compared to the frozen HLLM demonstrates that HLLM+ effectively addresses the trade-off between model performance and computational efficiency in recommendation systems.

Our initial findings demonstrated that a frozen item representation is adequate for recommendation systems. This opens the possibility of using other Language Models (LLMs), including multi-modal Visual Language Models (VLMs), which can generate embeddings from both images and text descriptions of items. To test this hypothesis, we conducted additional experiments, replacing the text-only model with the Smol VLM variant (HuggingFaceTB/SmolVLM-256M-Instruct [16]). The results of this evaluation are shown in Table 3 (train batch size 4, MAX TEXT LENGTH 128, MAX ITEM LIST LENGTH 5).

Model	LLM Architecture		Metrics (R)		Metrics (NCE)	
	Item	User	@10	@50	@10	@50
HLLM+	smol-t	smol-t	27.53	60.91	14.67	21.71
HLLM+	smol-m	smol-t	28.74	75.15	11.35	21.75

Table 3. HLLM+ can be adapted to use multimodal embedding for item representations without tuning of this model. Smol-t is a text only smol_llama101MGQA and smol-m is a multimodal SmolVLM-256M-Instruct.

The outcomes reveal that embeddings derived from VLMs can be effectively utilized within the user LLM and aligned using our proposed HLLM+ framework. This demonstrates the potential of our approach to incorporate various modalities, such as images and audio, in recommendation systems, leading to potentially higher system accuracy. As presented, multimodal embeddings can lead to higher recall, showing potential for improved performance. On the other hand, lower NCDG at 10 indicates that with fewer items the order might not be correct when using multimodal embeddings, and further experimentation is needed with multimodal datasets.

3.2. Ablation Study

To gain deeper insights into the effectiveness of HLLM+ and its individual components, we conduct a comprehensive ablation study. This analysis aims to isolate and evaluate the impact of key elements in our model architecture to determine long-tail recommendation performance, the effect

of frozen representations, the role of the projection layer, and the contribution of multi-head cross attention. The systematic examination of these components reveals their individual and collective contributions, highlighting HLLM+'s strengths and areas for potential optimization.

3.2.1. Long-tail Recommendation

While previous results demonstrate strong performance on popular items, crucial for mass-market campaigns and immediate advertising impact, we examine how the attention mechanism affects long-tail recommendations. Since attention weights tend to concentrate on frequently occurring items, this could potentially degrade performance for less common items, which may be needed for personalization and niche market segments in advertising technology. To evaluate this aspect, we extend our analysis to larger recommendation sets using R@50, and N@50 metrics, maintaining the same model configurations as in the previous experiments. Results of this evaluation are presented in Table 4.

Model	LLM Architecture		Metrics (R)		Metrics (NCE)	
	Item	User	@50	@100	@50	@100
HLLM	All	All	86.39	95.51	37.46	38.95
HLLM	Frozen	2 top	60.91	81.02	21.91	25.16
HLLM+	Frozen	2 top	71.64	88.90	26.58	29.38

Table 4. Long-tail recommendation performance comparison

The long-tail recommendation results reveal insightful performance differences across model configurations. As anticipated, the fully tunable HLLM demonstrates superior performance (R@50: 86.39, N@50: 37.46), showcasing its adaptability to less frequent items. The frozen HLLM, with limited tunability, shows a significant performance drop (R@50: 60.91, N@50: 21.91), highlighting the challenges in capturing nuanced long-tail relationships. Notably, HLLM+ strikes a balance, achieving improved performance (R@50: 71.64, N@50: 26.58) over the frozen HLLM despite using frozen item embeddings. While not matching the fully tunable model, HLLM+ demonstrates that its projection and cross-attention mechanisms effectively enhance long-tail recommendation capabilities. This suggests a viable trade-off between computational efficiency and accuracy, particularly valuable in resource-constrained scenarios.

3.2.2. Multi-head Cross Attention

Given these findings on long-tail performance limitations, we explore whether enhancing the cross-attention mechanism could improve the model's ability to capture diverse item relationships. Specifically, we investigate the impact of incorporating multiple attention heads and a diversity loss for self-supervised regularization. This approach aims to enable the model to simultaneously capture different aspects of user-item interactions, potentially addressing the challenge of long-tail recommendations by encouraging attention heads to focus on different parts of the user's history. The original cross-attention formulation 17 is modified with a multi-head cross-attention (MCA) variant that also produces a diversity loss:

$$\mathbf{Z}u, \mathcal{L}\text{div} = \text{MCA}(\mathbf{P}_u, \mathbf{E}_u, \mathbf{M}_u) \tag{17}$$

The multi-head cross-attention mechanism enables the model to capture different aspects of the user's historical interactions simultaneously, with queries derived from user representations and keys/values from historical items:

$$\mathbf{Q} = \mathbf{P}_u\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{E}_u\mathbf{W}^K, \quad \mathbf{V} = \mathbf{E}_u\mathbf{W}^V \tag{18}$$

where $\mathbf{P}_u$ is the projected user representation and $\mathbf{E}_u$ are the embeddings of the user's historical items. The attention computation proceeds through multiple heads:

$$\mathbf{A}_h = \text{softmax}\left(\frac{\mathbf{Q}_h\mathbf{K}_h^T}{\sqrt{d_h}}\right), \quad \mathbf{C}_h = \mathbf{A}_h\mathbf{V}_h \tag{19}$$

$$\mathbf{C} = \text{Concat}(\mathbf{C}_1, \ldots, \mathbf{C}_H)\mathbf{W}^O \tag{20}$$

$$\mathbf{Z}_u = \mathbf{P}_u + \sigma(g) \cdot \text{LayerNorm}(\mathbf{C}) \tag{21}$$

where $\mathbf{A}_h$ represents the attention weights for head h, $\mathbf{C}_h$ is the context vector for head h, H is the number of attention heads (4 by default), $d_h = d/H$ is the dimension per head, $\mathbf{W}^O$ is the output projection matrix, g is a learnable gate parameter initialized conservatively, and $\sigma(\cdot)$ is the sigmoid function.

The key innovation in this variant is the self-supervised diversity loss that encourages different attention heads to focus on different aspects of the user's history:

$$\bar{\mathbf{A}} = \frac{1}{H}\sum_{h=1}^{H}\mathbf{A}_h \tag{22}$$

$$\mathcal{L}_{\text{div}} = \frac{\lambda_{\text{div}}}{\text{Var}(\mathbf{A}_1, ..., \mathbf{A}_H) + \epsilon} \tag{23}$$

where $\text{Var}(\cdot)$ computes the variance across attention heads, λ_{div} is a diversity factor, and ϵ is a small constant for numerical stability. This diversity loss functions as a self-supervised regularization mechanism by penalizing redundancy across attention heads. When all heads attend to the same parts of the sequence, the variance is low and the loss is high; conversely, when heads attend to different parts, the variance increases and the loss decreases. This encourages

the model to develop specialized attention patterns across heads without requiring additional supervision. Results of MCA influence are presented in Table 5.

Model	R@5	R@10	R50	N@5	N@10	N@50
HLLM+CA	22.73	34.20	71.64	14.75	18.36	26.57
HLLM+MCA	18.42	29.00	64.86	11.82	15.23	23.01

Table 5. Performance comparison of HLLM+ with Cross Attention (CA) vs Multi-head Cross Attention (MCA)

Contrary to our hypothesis, the multi-head cross attention (MCA) mechanism did not yield performance improvements. As shown in Table 3, HLLM+MCA consistently underperformed compared to the simpler single-head cross attention model across all metrics, with decreases of 18.9% in R@5 and 19.8% in N@5. These results suggest that the added complexity of multiple attention heads and diversity loss may have overcomplicated the model architecture without providing additional benefit for capturing user-item relationships. The simpler cross attention mechanism appears to be sufficient for effectively modeling these interactions.

3.2.3. Impact of Layer Freezing Strategies

Then, we investigate the impact of different model freezing strategies. By systematically varying which components of the LLMs remain learnable or frozen for both user and item representations, we aim to understand the trade-offs between model adaptability and computational efficiency. Table 6 shows results of this analysis.

Trainability		Metrics			
Item	User	R@5	R@10	N@5	N@10
Frozen	2 layers	22.73	34.20	14.75	18.36
2 layers	Frozen	20.13	31.99	12.84	16.65
Frozen	Frozen	10.83	17.07	6.65	8.66
2 layers	2 layers	20.42	32.50	12.90	16.79
all layers	all layers	18.22	30.77	11.36	15.40

Table 6. Performance comparison of HLLM+ with different freezing configurations

The experimental results demonstrate that the choice of freezing strategy significantly impacts model performance. Keeping item representations frozen while allowing the top two user layers to be trainable yields the best performance across all metrics. Surprisingly, configurations with trainable item layers resulted in lower performance, regardless of user layer trainability. This may be attributed to the fact that items are shared across multiple users, making their representations more sensitive to changes and potentially leading to inconsistencies when adapted. The fully frozen configuration, while stable, shows substantially degraded performance (R@5: 10.83 vs 22.73), highlighting the importance of some degree of model adaptability. These findings suggest that asymmetric fine-tuning—focusing on user representation adaptation while maintaining stable item embeddings—provides the optimal balance between model flexibility and stability.

3.2.4. Projection Layer

We also analyze the role of the projection layer in transforming representations for cross-attention compatibility. To this end, we conducted experiments comparing three configurations of HLLM+: (1) without a projection layer, (2) with a projection layer as described in Eq.2, and (3) with a smaller projection layer consisting of $P_{u,1}$ only.

Projection Configuration	R@5	R@10	N@5	N@10
No Projection	19.23	30.32	12.20	15.77
One-layer Projection (Eq.5)	17.67	27.16	11.20	14.25
Two-layer Projection (Eq.6)	22.73	34.20	14.75	18.36

Table 7. Performance comparison of HLLM+ with different projection layer configurations

As proved in the experimental evaluation, the use of the projection layer is crucial for proper alignment of features between both representations. Interesingly, no projection turned out to be better than too shallow projection layer, indicating a need for more complexity to capture meaningful user interactions.

4. Discussion

Our experimental results demonstrate the significant potential of cross-attention mechanisms in enhancing recommendation accuracy without extensive model retraining. The HLLM+ approach, leveraging carefully designed alignment techniques, successfully outperformed frozen HLLM on popular item recommendations while tuning only 2 layers of a single language model, instead of fully tuning both. This not only addresses computational and cost challenges but also opens new possibilities for efficient adaptation of pre-trained models in recommendation systems, potentially achieving similar interest as in other systems for document processing, virtual assistants or content creation [25].

While our initial hypothesis suggested that multi-head cross-attention (MCA) could improve representation quality and enhance long-tail recommendations, empirical results showed that simpler architectures performed better. This finding highlights the importance of balancing model complexity with practical performance, suggesting that the added computational overhead of MCA might not always justify its use in current setups. Exploring other MCA-based variants, such as MCA Convolutional Encoders

should be further explored in the future [4]. Similarly, the use of the projection layer is crucial, but there is no need for a more complex version than a single MLP variant, according to the current results, that need further exploration, as well. Verification of other more advanced techniques such as two-phase training paradigm: the pre-training phase and the instruction-tuning phase and concepts from the alignment networks, e.g., AlignGPT [30] are interesting future directions of the proposed solution.

An important finding of our study is the asymmetry in how LLMs should be adapted for recommendation tasks. The observation that items are shared across multiple users while user representations are unique led to our successful strategy of keeping item representations frozen while allowing user-specific adaptations. This challenges the conventional symmetric approach to model fine-tuning and opens new research directions for exploring optimal training procedures. This results should be further verified with more architectures and datasets to confirm the finding.

Our experiments, although limited by computational resources to smaller language models, provide valuable insights for practical applications. While larger models like HLLM-1B and HLLM-7B [3] demonstrate superior performance, our results with SmolLM show promising potential for utilizing smaller, specialized models in recommendation systems. This is particularly relevant for businesses with limited resources, suggesting an interesting direction for future research: developing and pre-training domain-specific small models that could be efficiently reused across various brands and platforms through the proposed alignment. To further address such needs, neural architecture search [22] or quantization [12] could be utilized to find optimal architectures allowing for lower latency and cost, while maintaining required recommendation accuracy.

In addition, we replaced the text-based item LLM with a multimodal Visual Language Model (VLM). This experiment demonstrated that multimodal embeddings can be successfully aligned with the user LLM for recommendation tasks, outperforming single modality model. Although we employed a VLM capable of processing both text and images, our use of the Movie Lens dataset, which contains only textual information, limited our ability to fully exploit the model's multimodal capabilities. However, this scenario presents an intriguing research direction: enhancing e-commerce datasets with visual content could allow us to leverage the full potential of multimodal item representations. By utilizing both textual and visual data, our approach could generate more comprehensive item embeddings, potentially leading to more nuanced and accurate recommendations. Future work in this area could explore how incorporating multiple modalities impacts recommendation quality and user experience in diverse e-commerce contexts [29]. Also, combining our approach with visual attention tracking [11] to create more nuanced user interest profiles, or integrating speaker recognition for smart home applications [26]. These multimodal extensions could potentially enhance personalization capabilities while maintaining efficiency through our approach of aligning frozen representations.

Although the presented results are promising, additional experiments are required to further confirm the findings. First, studies with larger language models and diverse datasets are needed. Second, further investigation of training procedures that optimize the asymmetric adaptation strategy could yield even more efficient and effective models. Third, exploration of multimodal extensions for enhanced personalization could open new frontiers in recommendation systems. Lastly, the development of specialized small models for specific recommendation domains could democratize access to advanced AI-driven recommendations for a broader range of businesses [2].

These directions are particularly relevant for the evolving Advertisement and Marketing landscape, where balancing personalization capabilities with computational and cost efficiency becomes increasingly crucial for maintaining competitive advantage. By addressing these challenges, future research can build on our findings to create more powerful, efficient, and adaptable recommendation systems that meet the growing demands of the digital advertising market.

5. Conclusion

This work introduced HLLM+, demonstrating that effective recommendation systems can be built using frozen LLMs through careful representation alignment rather than extensive fine-tuning. Our approach achieved superior performance on popular item recommendations while keeping the item LLM completely frozen, suggesting a more efficient path forward for leveraging pre-trained language models in recommendation tasks. The success of asymmetric adaptation—focusing on user representation while maintaining stable item embeddings—challenges conventional assumptions about model fine-tuning in this domain. This architecture not only reduces computational requirements but also opens possibilities for flexible integration of various pre-trained models as item encoders.

Future work could explore: (1) dynamic adaptation strategies that balance model flexibility with computational efficiency, (2) enhanced techniques for long-tail item recommendations, and (3) extension to multi-modal recommendation scenarios leveraging frozen vision-language models. Additionally, investigating the application of our alignment approach to other recommendation scenarios and larger language models could further validate its effectiveness across different scales and domains.

References

[1] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Leandro von Werra, and Thomas Wolf. Smollm-blazingly fast and remarkably powerful, 2024. *URL: https://huggingface.co/blog/smollm*, 2024. 4

[2] Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin, and Pavlo Molchanov. Small language models are the future of agentic ai. *arXiv preprint arXiv:2506.02153*, 2025. 8

[3] Junyi Chen, Lu Chi, Bingyue Peng, and Zehuan Yuan. Hllm: Enhancing sequential recommendations via hierarchical large language models for item and user modeling. *arXiv preprint arXiv:2409.12740*, 2024. 2, 8

[4] Yi-Hui Chen, Eric Jui-Lin Lu, and Kwan-Ho Cheng. Integrating multi-head convolutional encoders with cross-attention for improved sparql query translation. *arXiv preprint arXiv:2408.13432*, 2024. 8

[5] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015. 4

[6] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*, 2015. 1

[7] Parshuram Hotkar, Rajiv Garg, and Kristen Sussman. Strategic social media marketing: An empirical analysis of sequential advertising. *Production and Operations Management*, 32 (12):4005–4020, 2023. 1

[8] IMARC Group. Advertising market size and share: Global industry trends, growth and forecast 2024-2033, 2024. Accessed: June 17, 2025. 1

[9] Jian Jia, Yipei Wang, Yan Li, Honggang Chen, Xuehan Bai, Zhaocheng Liu, Jian Liang, Quan Chen, Han Li, Peng Jiang, et al. Learn: Knowledge adaptation from large language model to recommendation for practical industrial application. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11861–11869, 2025. 1

[10] Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 197–206. IEEE, 2018. 1

[11] Alicja Kwaśniewska, Jacek Rumiński, and Paul Rad. Deep features class activation map for thermal face detection and tracking. In *2017 10Th international conference on human system interactions (HSI)*, pages 41–47. IEEE, 2017. 8

[12] Alicja Kwasniewska, Maciej Szankin, Mateusz Ozga, Jason Wolfe, Arun Das, Adam Zajac, Jacek Ruminski, and Paul Rad. Deep learning optimization for edge devices: Analysis of training quantization parameters. In *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*, pages 96–101. IEEE, 2019. 8

[13] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. Llara: Large language-recommendation assistant. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1785–1795, 2024. 1

[14] Jiahao Liu, Xueshuo Yan, Dongsheng Li, Guangping Zhang, Hansu Gu, Peng Zhang, Tun Lu, Li Shang, and Ning Gu. Improving llm-powered recommendations with personalized information. *arXiv preprint arXiv:2502.13845*, 2025. 1

[15] Yuanxing Liu, Wei-Nan Zhang, Yifan Chen, Yuchi Zhang, Haopeng Bai, Fan Feng, Hengbin Cui, Yongbin Li, and Wanxiang Che. Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue. *arXiv preprint arXiv:2310.14626*, 2023. 1

[16] Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025. 5

[17] Rajhans Mishra, Pradeep Kumar, and Bharat Bhasker. A web recommendation system considering sequential information. *Decision Support Systems*, 75:1–10, 2015. 1

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[19] Peter Szemraj and Vincent Haines. smol_llama101mgqa (revision 9c9c090), 2023. Accessed: June 17, 2025. 5

[20] Xubin Ren, Wei Wei, Lianghao Xia, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. Representation learning with large language models for recommendation. In *Proceedings of the ACM Web Conference 2024*, pages 3464–3475, 2024. 1

[21] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 811–820, 2010. 1

[22] Anthony Sarah, Sharath Nittur Sridhar, Maciej Szankin, and Sairam Sundaresan. Llama-nas: Efficient neural architecture search for large language models. In *European Conference on Computer Vision*, pages 67–74. Springer, 2025. 8

[23] Kowei Shih, Yi Han, and Li Tan. Recommendation system in advertising and streaming media: Unsupervised data enhancement sequence suggestions. *arXiv preprint arXiv:2504.08740*, 2025. 1

[24] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1441–1450, 2019. 1

[25] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. Pre-trained language models and their applications. *Engineering*, 25:51–65, 2023. 7

[26] Mingshan Wang, Tejaswini Sirlapu, Alicja Kwasniewska, Maciej Szankin, Marko Bartscherer, and Rey Nicolas. Speaker recognition using convolutional neural network with minimal training data for smart home solutions. In *2018 11th International Conference on Human System Interaction (HSI)*, pages 139–145. IEEE, 2018. 8

[27] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1929–1937, 2022. 1

[28] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, pages 346–353, 2019. 1

[29] Baixuan Xu, Weiqi Wang, Haochen Shi, Wenxuan Ding, Huihao Jing, Tianqing Fang, Jiaxin Bai, Xin Liu, Changlong Yu, Zheng Li, et al. Mind: Multimodal shopping intention distillation from large vision-language models for e-commerce purchase understanding. *arXiv preprint arXiv:2406.10701*, 2024. 8

[30] Fei Zhao, Taotian Pang, Chunhui Li, Zhen Wu, Junjie Guo, Shangyu Xing, and Xinyu Dai. Aligngpt: Multi-modal large language models with adaptive alignment capability. *arXiv preprint arXiv:2405.14129*, 2024. 8

[31] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. Deep interest evolution network for click-through rate prediction.(2018). *arXiv preprint arXiv:1809.03672*, 2018. 1