How Does Multimodal Training Affect Text-Only Recommendation Capabilities of LLMs: A Comparative Analysis

Mert Atay
Middle East Technical University
Ankara, Turkey
mert.atay@metu.edu.tr

Ismail Hakki Toroslu
Middle East Technical University
Ankara, Turkey
toroslu@ceng.metu.edu.tr

Ismail Sengor Altingovde
Middle East Technical University
Ankara, Turkey
altingovde@ceng.metu.edu.tr

Pinar Karagoz
Middle East Technical University
Ankara, Turkey
karagoz@ceng.metu.edu.tr

Abstract

Multimodal Large Language Models (MLLMs) appear as a promising future direction for enhancing Recommendation Systems (RSs), as they combine both efforts of integrating Large Language Models (LLMs) and multimodal data (e.g., images, video, and audio) into the RS domain. But can they fully replace LLMs, particularly in text-only recommendation tasks? Do MLLMs retain or even improve their textual recommendation capabilities? Since most MLLMs are not evaluated on text-only benchmarks, the question remains unanswered. In this work, text-only recommendation capabilities of five MLLMs with different architectures are investigated, along with their underlying LLM counterparts, using list-wise ranking task in four domains and three test parameters. Results show that despite undergoing multimodal training, MLLMs achieve comparable results to their underlying LLM counterparts in text-only recommendation, and outperform them in Movie domain, indicating that multimodal training has the potential to improve textual recommendation capabilities of LLMs.

CCS Concepts

• **Information systems** → **Recommender systems**.

Keywords

multimodal large language models, large language models, recommender systems, text-only recommendation, evaluation, list-wise ranking

ACM Reference Format:
Mert Atay, Ismail Hakki Toroslu, Ismail Sengor Altingovde, and Pinar Karagoz. 2025. How Does Multimodal Training Affect Text-Only Recommendation Capabilities of LLMs: A Comparative Analysis. In *Proceedings of The 1st Workshop on Next Generation of IR and Recommender Systems with Language Agents, Generative Models, and Conversational AI. (GENNEXT@SIGIR'25).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

GENNEXT@SIGIR'25, Padova, Italy

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recommender systems (RSs) play an important role in allowing users to discover relevant content based on their preferences, habits, and needs. With advancements in deep learning and computational resources, RSs have developed from simpler models and techniques to more complex and diverse technologies.

Following the advancements in pre-trained Large Language Models (LLMs), there is a growing interest in RS domain to use LLMs for recommendation by representing recommendation tasks as language modeling [3, 13, 21] and employing a wide range of methodologies [11, 27]. Efforts to improve RSs are not limited to LLM integration. As social and digital media platforms become more popular, available data and content now encompass a variety of modalities beyond text, including images, videos, and audio. Integration of these various modalities can improve RSs' capabilities and there are increasing efforts to include them in recommendation tasks [14, 15]. Hence, the literature points out the promising potential and the need for further exploration of employing new Multimodal Large Language Models (MLLMs) in RS.

MLLMs are LLM-based models that can process, reason with, and output multimodal data, along with text [30]. In the literature, MLLMs use divergent architectural designs and training methodologies [10]. In this work, the main focus is on MLLMs working with visual (image and/or video) modalities.

With their enhanced capabilities, MLLMs appear to be the next step in the evolution of LLMs. With data now spanning a variety of modalities, RS domain could benefit from employing MLLMs, but can they fully replace LLMs, particularly in text-only tasks? Since most MLLMs are not evaluated on text-only benchmarks, the question remains unanswered.

This question ultimately applies to the text-only recommendation task. In this work, the effects of multimodal training on text-only recommendation capabilities of LLMs are investigated with a comparative analysis of various MLLMs and their underlying LLM counterparts. To this aim, a variety of architectural designs [10] is considered. List-wise ranking in four different domains is employed as the recommendation task.

Table 1: Details and comparison of selected datasets across different domains (Book, Movie, Music, and News respectively).

Dataset	# Users	# Items	# Interactions	Sparsity	Interaction Type	Timestamp	User Metadata	Item Metadata
Amazon (2014) Books	8,026,324	2,370,585	22,507,155	99.99%	Ratings [0,5]	✓	✗	✓
MovieLens-1M	6,040	3,952	1,000,209	95.81%	Ratings [1,5]	✓	✓	✓
Amazon (2014) CDs & Vinyl	1,578,597	492,799	3,749,004	99.99%	Ratings [0,5]	✓	✗	✓
MIND-small (dev)	50,000	42,416	73,152	99.99%	Clicks {0,1}	✓	✗	✓

Table 2: Selected models for evaluating the impact of multimodal training on the text-only recommendation capabilities of LLMs.

MLLM	Underlying LLM Counterpart
Idefics3-8B-Llama3 [10]	Llama-3.1-8B-Instruct [4]
Qwen2-VL-7B-Instruct [24]	Qwen2-7B-Instruct [28]
Llama-3.2-11B-Vision-Instruct [4]	Llama-3.1-8B-Instruct [4]
Llava-v1.6-vicuna-7b-hf [12]	Vicuna-7b-v1.5 [31]
Llava-v1.6-mistral-7b-hf [12]	Mistral-7B-Instruct-v0.2 [8]

The main contribution of this work can be summarized as follows: 1) Text-only recommendation capabilities of MLLMs are investigated by presenting a comprehensive comparison with their underlying LLM counterparts. 2) Comparative analysis includes evaluation with varying test parameters across different domains, and a model catalog covering various architectural designs.

2 Related Work

Employing LLMs in RS domain starts with representing the recommendation task as language modeling [3, 13, 21]. This reformulation allows the use of strong zero-shot and few-shot capabilities of generative LLMs through prompting and in-context learning, which can be applied to recommendation domain. In that case, prompt construction is key to achieving optimal results, and recommendation capabilities of generative LLMs using various prompting techniques [1, 2, 6, 7, 19, 22], and in-context learning [7, 23, 25] are active areas of research, with efforts made to explore them.

Most recently, MLLMs have begun to be employed in RS for various tasks such as re-ranking [17] and sequential recommendation [29] with prompting and in-context learning using multimodal data. Several domain-specific applications also exist, such as [9] for e-commerce and [20] for personalized multimodal generation. Additionally, survey efforts highlight the application of multimodal pre-training, adaptation, and generation in the RS domain [16].

However, to the best of our knowledge, the evaluation of MLLMs' text-only recommendation capabilities in RS remains unexplored in the literature.

3 Methodology

Problem Definition. In this work, pre-trained LLMs and MLLMs are used as RSs on list-wise ranking task via prompting and in-context learning.

Parametric prompts (e.g., Figure 1) are used, which can be adapted to different domains based on different parameters (varying number of demonstration examples [zero-shot and few-shot], history items and candidate items). During the creation of prompts, prominent

```
You are a movie recommender system. Your task is to rank a list of
candidate movies based on a user's watching history, from most preferred
to least preferred. The movie that is most likely to be preferred should be
ranked first (index 0), the second-most preferred should be ranked second
(index 1), and so on, with the least preferred ranked last.
You will be provided with example rankings, enclosed in triple backticks
(“‘), to help guide your decision-making.
Your output should be a sequence of number indexes (e.g., 0, 1, 2, 3, 4, ...)
corresponding to the order of movie preferences, starting with the most
preferred. Do not include any explanations or information.

“‘{demonstration_examples}“‘

User History:
{user_history}
Candidate movies:
0. {candidate_item1}
1. {candidate_item2}
...
(n-2). {candidate_item(n-1)}
(n-1). {candidate_item(n)}
Output:
```

Figure 1: Prompt template in movie domain for list-wise ranking.

prompt engineering guides[1] [2] are followed. Formally, given a set of users U and a set of items I, for a given user $u \in U$, the following are defined as follows: a set of k candidate items $c_u = \{i_1, i_2, \ldots, i_k\} \subset I$, a sequence of n historical interactions $h_u = [i'_1, i'_2, \ldots, i'_n] \subset I$, and a sequence of predictions $y_u = [y_1, y_2, \ldots, y_k] \subset I$, where y_u is a permutation of c_u and $|y_u| = |c_u| = k$. In the case of this work, an item $i \in I$ is a string, corresponding to the item's title.

Let E be a subset of N randomly chosen users from U such that $E \subset U$ and $|E| = N$. Demonstration examples D are compiled as: $D = \emptyset$ if $N = 0$; $\{f(h_u, c_u, y_u) : u \in E\}$ if $N \geq 1$

A random group of N users are selected from U. For each selected user u, demonstration examples D are compiled by taking their h_u, c_u and y_u (ground truth ranking). These details are processed into a formatted string that fits into the prompt template. If $N = 0$ (zero-shot), no examples are included in the prompt template ($D = \emptyset$).

The query for a given user u', where $u' \in U \setminus E$, is created by the function g, which constructs the final prompt based on a specific domain. Given a domain $dmn \in \{\text{“}Movie\text{”}, \text{“}Book\text{”}, \text{“}Music\text{”}, \text{“}News\text{”}\}$, the set of demonstration examples D, the user's historical interactions and the candidate items: $prompt = g_{dmn}(D, h_{u'}, c_{u'})$. Finally,

[1] https://platform.openai.com/docs/guides/prompt-engineering
[2] https://cloud.google.com/discover/what-is-prompt-engineering

Table 3: Evaluation of full model catalog in list-wise ranking task with default parameters ($n_{shot} = 1$, $h_{size} = 5$, $c_{size} = 5$). The best and the second best results are highlighted for each domain. Non-compliant responses are penalized.

Domain	Model	Model Type	NDCG@3	MRR@3	Compliance Rate(%)
Book	Idefics3-8B-Llama3	MLLM	0.4364 ± 0.0064	0.3794 ± 0.0057	98.0667 ± 1.3317
	Qwen2-VL-7B-Instruct	MLLM	**0.4647 ± 0.0012**	**0.4077 ± 0.0012**	99.8000 ± 0.0000
	Llama-3.2-11B-Vision-Instruct	MLLM	0.0103 ± 0.0028	0.0090 ± 0.0024	2.0667 ± 0.6110
	Llava-v1.6-vicuna-7b-hf	MLLM	0.4191 ± 0.0070	0.3629 ± 0.0056	99.8667 ± 0.1155
	Llava-v1.6-mistral-7b-hf	MLLM	0.3504 ± 0.1218	0.3002 ± 0.1051	79.6000 ± 26.5458
	Llama-3.1-8B-Instruct	LLM	0.3305 ± 0.0338	0.2848 ± 0.0295	73.8667 ± 7.9053
	Qwen2-7B-Instruct	LLM	0.4560 ± 0.0155	0.3955 ± 0.0137	98.5333 ± 2.0429
	Vicuna-7b-v1.5	LLM	0.4225 ± 0.0124	0.3639 ± 0.0097	97.8667 ± 0.9018
	Mistral-7B-Instruct-v0.2	LLM	0.1680 ± 0.0680	0.1445 ± 0.0590	39.7333 ± 16.0179
Movie	Idefics3-8B-Llama3	MLLM	**0.4922 ± 0.0014**	**0.4354 ± 0.0013**	99.7333 ± 0.1155
	Qwen2-VL-7B-Instruct	MLLM	0.4445 ± 0.0113	0.3869 ± 0.0104	99.4000 ± 0.0000
	Llama-3.2-11B-Vision-Instruct	MLLM	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	Llava-v1.6-vicuna-7b-hf	MLLM	0.3213 ± 0.1033	0.2799 ± 0.0920	76.1333 ± 25.3001
	Llava-v1.6-mistral-7b-hf	MLLM	0.3361 ± 0.0892	0.2900 ± 0.0760	76.2000 ± 19.9570
	Llama-3.1-8B-Instruct	LLM	0.0437 ± 0.0280	0.0366 ± 0.0244	8.9333 ± 6.2365
	Qwen2-7B-Instruct	LLM	0.4845 ± 0.0029	0.4217 ± 0.0020	98.7333 ± 0.3055
	Vicuna-7b-v1.5	LLM	0.4377 ± 0.0107	0.3800 ± 0.0098	98.5333 ± 1.3317
	Mistral-7B-Instruct-v0.2	LLM	0.2444 ± 0.0945	0.2117 ± 0.0823	55.0000 ± 21.5527
Music	Idefics3-8B-Llama3	MLLM	0.3804 ± 0.0466	0.3300 ± 0.0386	84.0000 ± 11.2303
	Qwen2-VL-7B-Instruct	MLLM	**0.4335 ± 0.0072**	**0.3772 ± 0.0062**	98.9333 ± 0.6110
	Llama-3.2-11B-Vision-Instruct	MLLM	0.0070 ± 0.0016	0.0060 ± 0.0015	1.6667 ± 0.3055
	Llava-v1.6-vicuna-7b-hf	MLLM	0.3548 ± 0.0240	0.3038 ± 0.0215	85.6667 ± 6.0044
	Llava-v1.6-mistral-7b-hf	MLLM	0.1939 ± 0.0528	0.1662 ± 0.0476	45.9333 ± 11.2006
	Llama-3.1-8B-Instruct	LLM	0.1541 ± 0.0476	0.1308 ± 0.0403	36.6667 ± 10.6308
	Qwen2-7B-Instruct	LLM	0.4307 ± 0.0229	0.3737 ± 0.0209	94.4667 ± 2.7006
	Vicuna-7b-v1.5	LLM	0.3934 ± 0.0084	0.3385 ± 0.0069	90.6000 ± 2.2271
	Mistral-7B-Instruct-v0.2	LLM	0.0728 ± 0.0149	0.0650 ± 0.0133	14.6000 ± 3.0000
News	Idefics3-8B-Llama3	MLLM	0.4337 ± 0.0067	0.3753 ± 0.0052	95.6667 ± 5.4271
	Qwen2-VL-7B-Instruct	MLLM	0.4182 ± 0.0081	0.3634 ± 0.0038	98.6000 ± 0.7211
	Llama-3.2-11B-Vision-Instruct	MLLM	0.0877 ± 0.0232	0.0759 ± 0.0192	19.0000 ± 4.6130
	Llava-v1.6-vicuna-7b-hf	MLLM	0.4173 ± 0.0035	0.3574 ± 0.0032	99.9333 ± 0.1155
	Llava-v1.6-mistral-7b-hf	MLLM	0.2337 ± 0.0951	0.2014 ± 0.0827	50.9333 ± 20.9774
	Llama-3.1-8B-Instruct	LLM	0.3671 ± 0.0339	0.3179 ± 0.0304	83.4000 ± 9.8000
	Qwen2-7B-Instruct	LLM	**0.4481 ± 0.0072**	**0.3894 ± 0.0060**	98.5333 ± 0.7024
	Vicuna-7b-v1.5	LLM	0.4240 ± 0.0069	0.3623 ± 0.0046	99.1333 ± 0.4619
	Mistral-7B-Instruct-v0.2	LLM	0.1712 ± 0.0179	0.1462 ± 0.0141	41.8000 ± 6.4715

the predictions which represent the list-wise ranking of the candidate items $c_{u'}$ for the given user u', are generated by inputting *prompt* into LLMs and MLLMs. Same prompts are input to both model types. The predictions are then evaluated using ranking evaluation metrics.

Model Catalog. Model selection (Table 2) involves MLLMs with diverse architectural designs and training strategies [10], along with their underlying LLM counterparts. Hugging Face's Model Hub[3] and transformers library[4] are used to work locally with the models.

[3]https://huggingface.co/models
[4]https://huggingface.co/docs/transformers/en/index

Datasets. Datasets from four different domains are selected to allow an evaluation in diverse contexts: 'Book' subset of the Amazon Product Reviews (2014) dataset [18] from Book, MovieLens-1M dataset [5] from Movie, 'CDs & Vinyl' subset of the same Amazon Product Reviews (2014) dataset [18] from Music, and the validation set of the MIND-Small dataset [26] from News domains. The details of each dataset are given in Table 1. For Movie, Book, and Music datasets, items with ratings of 4 or higher (out of 5) are considered as positive, while those with lower ratings as negative. For News dataset, the original binary labels are used. As a pre-processing step, items with no title information are discarded since the titles are used during the prompt generation.

Following the formal problem definition, for a user $u \in U$, the full history sequence H_u is defined as the sequence of positive items the user has interacted with, ordered by recency. The historical

Table 4: Best performing n_{shot} values for each domain and selected model. The best and the second best results are highlighted for each domain. Non-compliant responses are penalized.

Domain	Model	Model Type	n_{shot}	NDCG@3	MRR@3	Compliance Rate(%)
Book	Idefics3-8B-Llama3	MLLM	2	0.4525 ± 0.0183	0.3974 ± 0.0172	99.6667 ± 0.2309
	Qwen2-VL-7B-Instruct	MLLM	5	**0.4807 ± 0.0076**	**0.4253 ± 0.0067**	99.9333 ± 0.1155
	Llama-3.1-8B-Instruct	LLM	5	0.4443 ± 0.0126	0.3820 ± 0.0124	99.8667 ± 0.2309
	Qwen2-7B-Instruct	LLM	0	0.4774 ± 0.0000	0.4237 ± 0.0000	96.6000 ± 0.0000
Movie	Idefics3-8B-Llama3	MLLM	2	**0.5121 ± 0.0114**	**0.4540 ± 0.0094**	99.9333 ± 0.1155
	Qwen2-VL-7B-Instruct	MLLM	0	0.4819 ± 0.0000	0.4240 ± 0.0000	98.4000 ± 0.0000
	Llama-3.1-8B-Instruct	LLM	5	0.4287 ± 0.0161	0.3668 ± 0.0134	97.2000 ± 2.1166
	Qwen2-7B-Instruct	LLM	4	0.4903 ± 0.0053	0.4327 ± 0.0069	98.6000 ± 0.2000
Music	Idefics3-8B-Llama3	MLLM	5	0.4459 ± 0.0050	0.3841 ± 0.0053	99.7333 ± 0.3055
	Qwen2-VL-7B-Instruct	MLLM	1	0.4358 ± 0.0085	0.3787 ± 0.0070	99.2667 ± 0.3055
	Llama-3.1-8B-Instruct	LLM	4	0.4384 ± 0.0028	0.3751 ± 0.0028	99.6667 ± 0.5774
	Qwen2-7B-Instruct	LLM	2	**0.4524 ± 0.0121**	**0.3941 ± 0.0120**	94.8667 ± 1.9009
News	Idefics3-8B-Llama3	MLLM	3	0.4719 ± 0.0074	0.4129 ± 0.0057	100.0000 ± 0.0000
	Qwen2-VL-7B-Instruct	MLLM	5	0.4412 ± 0.0035	0.3836 ± 0.0039	99.5333 ± 0.2309
	Llama-3.1-8B-Instruct	LLM	3	0.4464 ± 0.0142	0.3816 ± 0.0129	99.0000 ± 1.0583
	Qwen2-7B-Instruct	LLM	4	**0.4840 ± 0.0086**	**0.4264 ± 0.0104**	99.6000 ± 0.2000

Table 5: Best performing h_{size} values for each domain and selected model. The best and the second best results are highlighted for each domain. Non-compliant responses are penalized.

Domain	Model	Model Type	h_{size}	NDCG@3	MRR@3	Compliance Rate(%)
Book	Idefics3-8B-Llama3	MLLM	5	0.4357 ± 0.0244	0.3780 ± 0.0210	95.2667 ± 3.8018
	Qwen2-VL-7B-Instruct	MLLM	5	0.4707 ± 0.0070	0.4114 ± 0.0059	99.7333 ± 0.2309
	Llama-3.1-8B-Instruct	LLM	25	0.4321 ± 0.0116	0.3719 ± 0.0121	97.0000 ± 1.5620
	Qwen2-7B-Instruct	LLM	10	**0.4823 ± 0.0065**	**0.4242 ± 0.0076**	98.0000 ± 0.4000
Movie	Idefics3-8B-Llama3	MLLM	5	**0.5029 ± 0.0105**	**0.4444 ± 0.0070**	99.7333 ± 0.1155
	Qwen2-VL-7B-Instruct	MLLM	50	0.4553 ± 0.0060	0.3988 ± 0.0050	99.8667 ± 0.1155
	Llama-3.1-8B-Instruct	LLM	25	0.4320 ± 0.0046	0.3737 ± 0.0045	98.9333 ± 0.2309
	Qwen2-7B-Instruct	LLM	5	0.4838 ± 0.0008	0.4228 ± 0.0013	98.5333 ± 0.4619
Music	Idefics3-8B-Llama3	MLLM	50	0.4246 ± 0.0113	0.3630 ± 0.0097	98.2000 ± 1.0000
	Qwen2-VL-7B-Instruct	MLLM	5	**0.4318 ± 0.0125**	**0.3762 ± 0.0108**	98.8000 ± 0.3464
	Llama-3.1-8B-Instruct	LLM	25	0.4167 ± 0.0080	0.3592 ± 0.0068	98.0000 ± 1.8330
	Qwen2-7B-Instruct	LLM	5	**0.4318 ± 0.0094**	0.3749 ± 0.0071	95.2667 ± 1.3013
News	Idefics3-8B-Llama3	MLLM	10	0.4175 ± 0.0145	0.3587 ± 0.0093	99.5333 ± 0.3055
	Qwen2-VL-7B-Instruct	MLLM	5	0.4221 ± 0.0227	0.3665 ± 0.0177	99.3333 ± 0.6429
	Llama-3.1-8B-Instruct	LLM	50	0.4403 ± 0.0102	0.3851 ± 0.0077	92.7333 ± 5.6012
	Qwen2-7B-Instruct	LLM	10	**0.4648 ± 0.0069**	**0.4064 ± 0.0067**	98.7333 ± 0.1155

interactions sequence h_u is then formed by taking the most recent n items from H_u, excluding the most recent one which is reserved for the candidate set ($h_u \subset H_u$).

During the creation of the candidate item set c_u, the most recent item in H_u is taken as the single positive candidate, while the remaining $k-1$ candidates are randomly selected from the user's negative items, following the common practice [1, 17, 21]. Different subsets of the datasets are created to test varying historical interaction sequence and candidate item set sizes. The following conditions must hold during the creation of respective subsets: $|H_u| - 1 \geq n$ and $|Neg_u| \geq k-1$, where 1 corresponds to the single (most recent) positive item and Neg_u is the set of negative items the user has interacted with.

4 Experimental Setup & Results

500 data points[5] are sampled for each test case, as sampling is widely adopted in LLM-based recommender systems due to high computational costs associated with large-scale evaluations [1, 2, 7, 17]. Three test parameters are defined as follows: Number of Demonstration Examples (n_{shot}), Number of Historical Interactions (h_{size}), Number of Candidate Items (c_{size}). In the default scenario, the test parameters are set as: $n_{shot} = 1$, $h_{size} = 5$, $c_{size} = 5$.

Initially, the full model catalog is evaluated with the default scenario. Based on the results, the top performant four models are

[5]With the exception of 442 data points for Book Dataset's $c_{size} = 50$ case, as it is the maximum number of available data points that meets the subset creation criteria.

Table 6: Best results are achieved with $c_{size} = 5$ for each domain and selected model. The best and the second best results are highlighted for each domain. Non-compliant responses are penalized.

Domain	Model	Model Type	NDCG@3	MRR@3	Compliance Rate(%)
Book	Idefics3-8B-Llama3	MLLM	0.4173 ± 0.0136	0.3622 ± 0.0150	95.4667 ± 2.0817
	Qwen2-VL-7B-Instruct	MLLM	**0.4484 ± 0.0263**	**0.3891 ± 0.0284**	99.6667 ± 0.4163
	Llama-3.1-8B-Instruct	LLM	0.3350 ± 0.0411	0.2883 ± 0.0358	75.7333 ± 9.1528
	Qwen2-7B-Instruct	LLM	0.4450 ± 0.0256	0.3826 ± 0.0198	98.5333 ± 0.8083
Movie	Idefics3-8B-Llama3	MLLM	**0.4723 ± 0.0455**	**0.4140 ± 0.0462**	98.6667 ± 1.6166
	Qwen2-VL-7B-Instruct	MLLM	0.4412 ± 0.0153	0.3833 ± 0.0178	99.4000 ± 0.2000
	Llama-3.1-8B-Instruct	LLM	0.1012 ± 0.1141	0.0881 ± 0.1010	21.4000 ± 23.7394
	Qwen2-7B-Instruct	LLM	0.4636 ± 0.0288	0.4016 ± 0.0283	98.9333 ± 0.6110
Music	Idefics3-8B-Llama3	MLLM	0.4071 ± 0.0324	0.3539 ± 0.0331	92.3333 ± 3.9260
	Qwen2-VL-7B-Instruct	MLLM	**0.4288 ± 0.0110**	**0.3721 ± 0.0101**	98.2667 ± 0.9238
	Llama-3.1-8B-Instruct	LLM	0.1776 ± 0.0790	0.1521 ± 0.0677	40.8667 ± 19.8316
	Qwen2-7B-Instruct	LLM	0.4149 ± 0.0372	0.3613 ± 0.0327	93.1333 ± 5.8287
News	Idefics3-8B-Llama3	MLLM	0.4302 ± 0.0061	0.3727 ± 0.0042	99.4000 ± 0.5292
	Qwen2-VL-7B-Instruct	MLLM	0.4359 ± 0.0158	0.3756 ± 0.0178	99.6667 ± 0.1155
	Llama-3.1-8B-Instruct	LLM	0.4336 ± 0.0225	0.3732 ± 0.0162	99.8667 ± 0.2309
	Qwen2-7B-Instruct	LLM	**0.4421 ± 0.0085**	**0.3818 ± 0.0108**	98.8000 ± 0.4000

Table 7: Evaluation of selected models with best performing test parameter values for each domain. $c_{size} = 5$ gives the best result for all scenarios.The best and the second best results are highlighted for each domain. Non-compliant responses are penalized.

Domain	Model	Model Type	n_{shot}	h_{size}	NDCG@3	MRR@3	Compliance Rate(%)
Book	Idefics3-8B-Llama3	MLLM	2	5	0.4518 ± 0.0095	0.3970 ± 0.0085	99.2000 ± 0.2000
	Qwen2-VL-7B-Instruct	MLLM	5	5	**0.4757 ± 0.0034**	**0.4194 ± 0.0025**	99.5333 ± 0.4163
	Llama-3.1-8B-Instruct	LLM	5	25	0.4302 ± 0.0162	0.3712 ± 0.0156	99.3333 ± 1.1547
	Qwen2-7B-Instruct	LLM	0	10	0.2115 ± 0.0000	0.1853 ± 0.0000	47.6000 ± 0.0000
Movie	Idefics3-8B-Llama3	MLLM	2	5	**0.5051 ± 0.0180**	**0.4470 ± 0.0165**	100.0000 ± 0.0000
	Qwen2-VL-7B-Instruct	MLLM	0	50	0.4352 ± 0.0000	0.3760 ± 0.0000	99.6000 ± 0.0000
	Llama-3.1-8B-Instruct	LLM	5	25	0.4402 ± 0.0156	0.3824 ± 0.0130	99.9333 ± 0.1155
	Qwen2-7B-Instruct	LLM	4	5	0.4849 ± 0.0055	0.4280 ± 0.0060	98.5333 ± 0.1155
Music	Idefics3-8B-Llama3	MLLM	5	50	0.4168 ± 0.0436	0.3572 ± 0.0352	96.0667 ± 6.8127
	Qwen2-VL-7B-Instruct	MLLM	1	5	0.4337 ± 0.0075	0.3781 ± 0.0071	98.8667 ± 0.3055
	Llama-3.1-8B-Instruct	LLM	4	25	0.4313 ± 0.0049	0.3699 ± 0.0022	99.7333 ± 0.3055
	Qwen2-7B-Instruct	LLM	2	5	**0.4609 ± 0.0063**	**0.4019 ± 0.0070**	96.6667 ± 0.8083
News	Idefics3-8B-Llama3	MLLM	3	10	0.4236 ± 0.0078	0.3609 ± 0.0063	100.0000 ± 0.0000
	Qwen2-VL-7B-Instruct	MLLM	5	5	0.4452 ± 0.0047	0.3862 ± 0.0029	99.5333 ± 0.3055
	Llama-3.1-8B-Instruct	LLM	3	50	0.4186 ± 0.0527	0.3596 ± 0.0433	91.4000 ± 12.0216
	Qwen2-7B-Instruct	LLM	4	10	**0.4601 ± 0.0187**	**0.4014 ± 0.0180**	99.0667 ± 0.7024

selected (2 MLLMs and their underlying LLM counterparts) for further evaluation with varying test parameters. While testing a given parameter, the other two test parameters are set to their default values. The following values are used for the three test parameters: $n_{shot} = [0, 1, 2, 3, 4, 5]$, $h_{size} = [5, 10, 25, 50]$, $c_{size} = [5, 10, 25, 50]$. In order to obtain deterministic outputs from the models, greedy search decoding method is used during text generation. The results are reported in Normalized Discounted Cumulative Gain (NDCG@K) and Mean Reciprocal Rank (MRR@K) with $K = 3$.

There are instances where the models can generate responses that do not comply with the given prompt which asks models to output only the sequence of numerical indices representing the ranked order of item preferences, starting with the most preferred item and excluding any explanations or additional information. A response is considered non-compliant if it does not begin with a sequence of numerical indices, contains unintended explanations, or includes duplicated or missing indices. Punctuations are ignored. In this study, compliance with the prompt is considered as an additional indicator of a model's performance and compliance rates are calculated as the ratio of compliant responses to total number of test instances as given in Equation 1.

$$\text{Compliance Rate} = \frac{\text{Number of Compliant Answers}}{\text{Number of Test Examples}} \times 100 \quad (1)$$

In the other metrics, non-compliant responses are treated as false predictions. Each test case is repeated three times and the average

is reported to account for the variability introduced by the random ordering of candidate items during prompt generation.

Initially, the full model catalog is evaluated across four domains using the default parameter values to identify the top-performing models for further evaluation with test parameters. Based on the results, two best-performing MLLMs, Idefics3-8B-Llama3 and Qwen2-VL-7B-Instruct are selected, along with their respective LLM counterparts, Llama-3.1-8B-Instruct and Qwen2-7B-Instruct, for further evaluation.

The initial evaluation with the complete model catalog (Table 3) shows that in three of the four domains, MLLMs show higher performance. Idefics3-8B-LLama3 (MLLM) outperforms LLama-3.1-8B-Instruct (LLM) across all domains. For Llava-v1.6 models (MLLMs), Mistral-7B-Instruct-v0.2 (LLM) based model shows improvements, while Vicuna-7b-v1.5 (LLM) version shows declines, with the amount of improvement and decline varying across different domains. However, in both cases, Llava-v1.6 models fall behind other MLLMs and LLMs in performance. Qwen2-VL-7B-Instruct (MLLM) outperforms Qwen2-7B-Instruct (LLM) in Book and Music domains with small margins, but falls behind in Movie and News domains.

LLama-3.2-11B-Vision-Instruct (MLLM) shows an interesting case. Looking at Table 3, it is evident that it shows poor performance and low compliance rates, making it an ineffective RS. LLama-3.2-11B-Vision-Instruct model differs from other MLLMs in the model catalog in both its architectural design and its LLM training methodology. Considering the limitations of the test case scenario, which focuses on list-wise ranking only, it is suggested that the special case of LLama-3.2-11B-Vision-Instruct should be studied further in future research.

For the evaluations of test parameters (Tables 4, 5, 6, 7), all models achieve the best results when working with small candidate sizes ($c_{size} = 5$). Figures 2, 3, 4 (due to space considerations, excluding MRR@3 and Compliance Rates) show the details of the results for each test parameter with NDCG@3.

Table 7 shows the results for using optimal values for each test parameter. However, using these values together does not necessarily improve the performance (e.g., Table 4: Qwen2-VL-7B-Instruct [Book], Idefics3-8B-Llama3 [Movie]).

Looking at the bigger picture, it is evident that MLLMs, especially Idefics3-8B-LLama3 and Qwen2-VL-7B-Instruct perform comparably to their underlying LLM counterparts and, in the Movie domain, even outperform them. This finding may indicate that multimodal training has the potential to improve textual recommendation capabilities for specific domains. The results show that MLLMs are a promising direction for RS. In certain domains, in the case of this work Movie domain, they can even replace LLMs on text-only recommendation tasks.

Furthermore, in Figure 3, it is observed that MLLMs can achieve comparable performance to LLMs with smaller history sizes for some domains (e.g., Book domain).

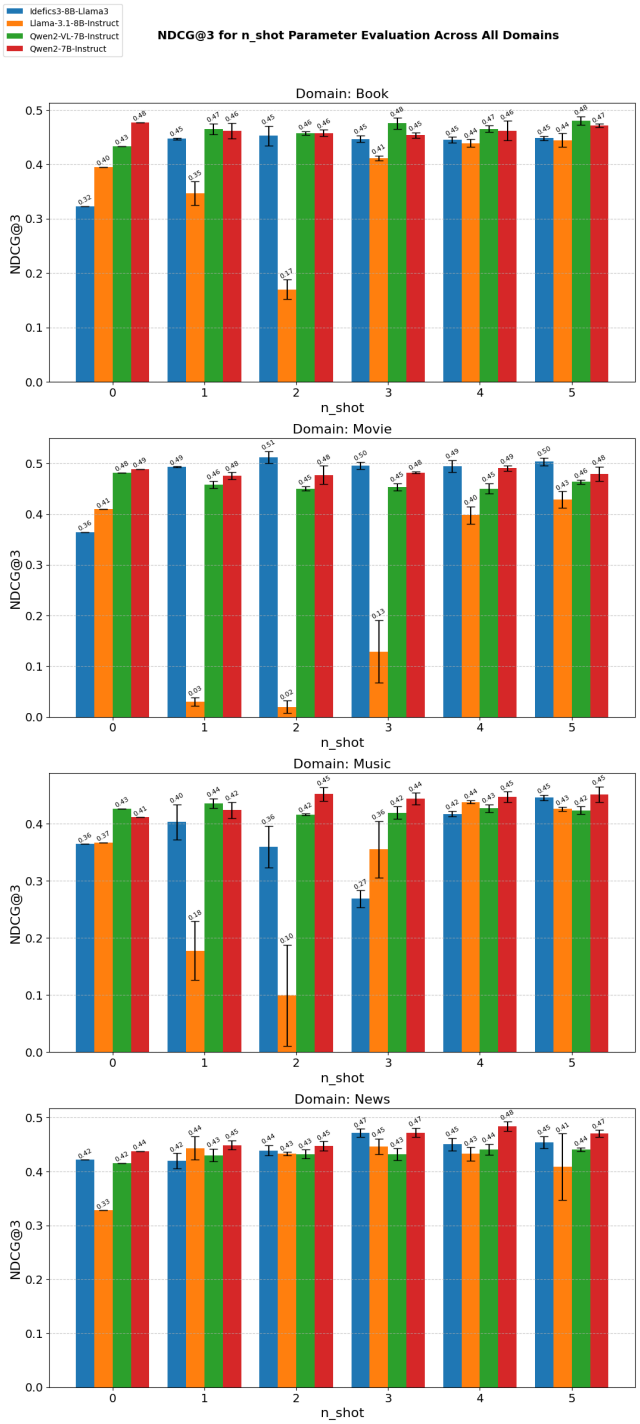

Figure 2: NDCG@3 values for n_{shot} parameter evaluation across four domains. Non-compliant responses are penalized.

5 Conclusion

In this study, text-only recommendation capabilities of five MLLMs with diverse architectural designs, along with their underlying LLM counterparts, are evaluated using list-wise ranking task across four

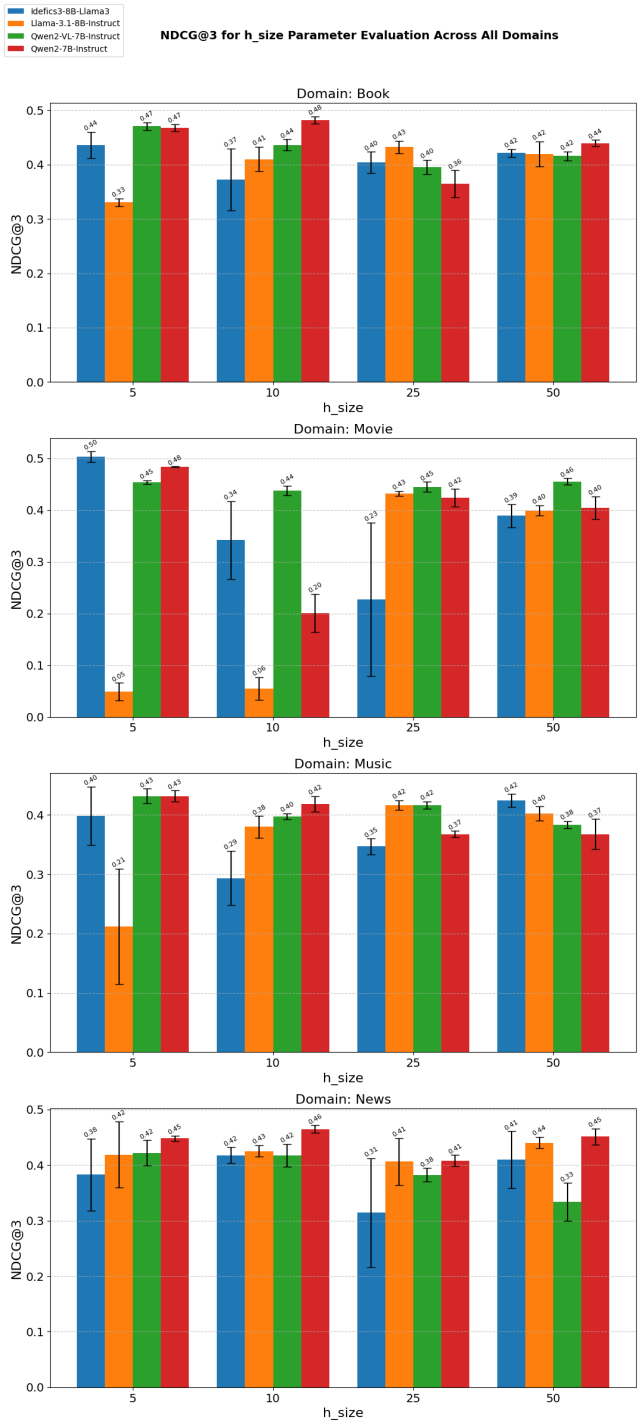

Figure 3: NDCG@3 values for h_{size} parameter evaluation across four domains. Non-compliant responses are penalized.

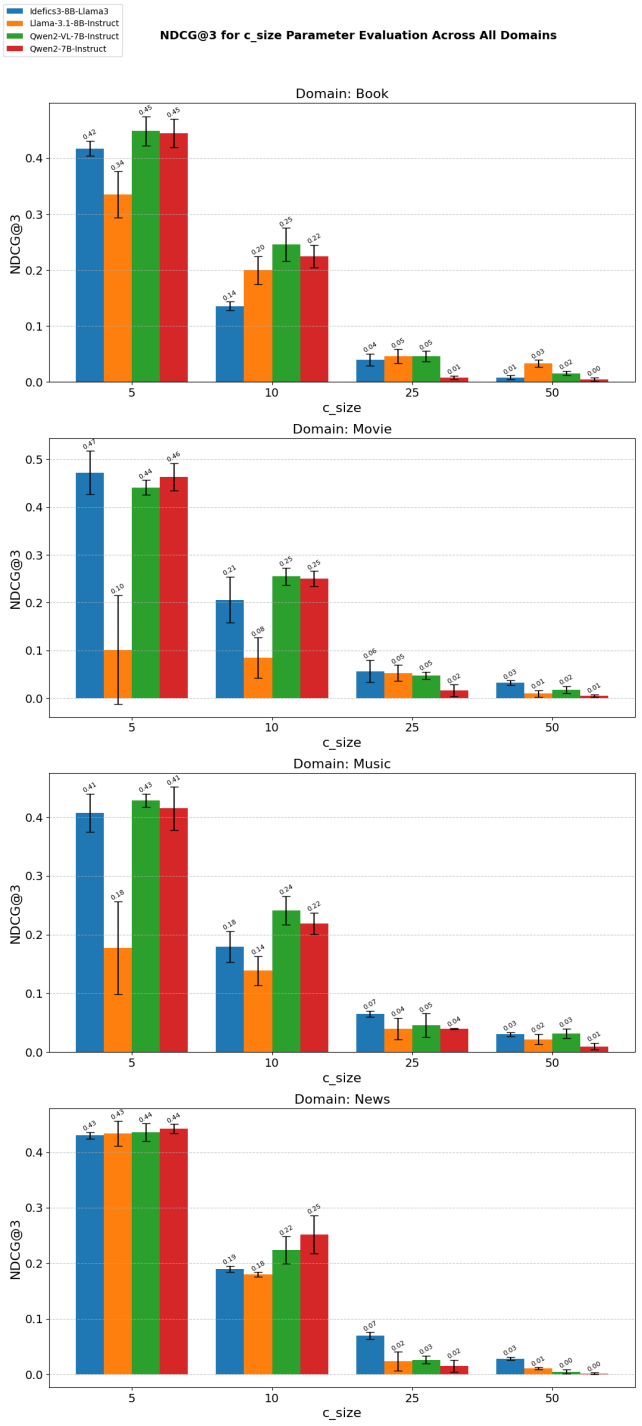

Figure 4: NDCG@3 values for c_{size} parameter evaluation across four domains. Non-compliant responses are penalized.

different domains and three test parameters. Experimental results show that all models perform better with small candidate sizes. Overall, it is evident that MLLMs perform comparably to their underlying LLM counterparts and, in the Movie domain, even outperform them. This finding may indicate that multimodal training has the potential to improve the textual recommendation capabilities of LLMs for specific domains, making MLLMs a promising direction for RS. In certain domains, Movie domain in this work, they can even replace LLMs on text-only recommendation tasks.

6 Limitations

Although this preliminary comparative analysis investigates the effect of multimodal training on text-only recommendation capabilities of LLMs, there are several limitations that should be noted.

Firstly, the recommendation task focused on list-wise ranking only, which may not fully capture all RS tasks. Future work should broaden the evaluation to include various recommendation tasks such as point-wise, pair-wise, or conversational recommendation.

Secondly, even though datasets from diverse domains are employed, the number of datasets could be increased to cover an even broader range of domains.

Another limitation is that the study focused on smaller models with approximately 7 billion parameters due to the high computational resources required for running larger models. As a result, the performance of larger models, which could potentially show better results, could not be assessed. Future work could explore the text-only recommendation capabilities of much larger models, with parameters ranging from 70 to 90 billion, to better highlight their potential.

Acknowledgments

This work is partially funded by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under grant no. 5230039 and METU under the grant no. ADEP-312-2024-11484

References

[1] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1126–1132. doi:10.1145/3604915.3610646

[2] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. arXiv:2303.14524 [cs.IR] https://arxiv.org/abs/2303.14524

[3] Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5). In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. Association for Computing Machinery, New York, NY, USA, 299–315. doi:10.1145/3523227.3546767

[4] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[5] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages. doi:10.1145/2827872

[6] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian Mcauley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 720–730. doi:10.1145/3583780.3614949

[7] Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large Language Models are Zero-Shot Rankers for Recommender Systems. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part II* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 364–381. doi:10.1007/978-3-031-56060-6_24

[8] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[9] Saketh Reddy Karra and Theja Tulabandhula. 2024. InteraRec: Interactive Recommendations Using Multimodal Large Language Models. In *Trends and Applications in Knowledge Discovery and Data Mining*, Zhaoxia Wang and Chang Wei Tan (Eds.). Springer Nature Singapore, Singapore, 32–43.

[10] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. 2024. Building and better understanding vision-language models: insights and future directions. arXiv:2408.12637 [cs.CV] https://arxiv.org/abs/2408.12637

[11] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2025. How Can Recommender Systems Benefit from Large Language Models: A Survey. *ACM Trans. Inf. Syst.* 43, 2, Article 28 (Jan. 2025), 47 pages. doi:10.1145/3678004

[12] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/

[13] Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2023. Pre-train, Prompt, and Recommendation: A Comprehensive Survey of Language Modeling Paradigm Adaptations in Recommender Systems. *Transactions of the Association for Computational Linguistics* 11 (2023), 1553–1571. doi:10.1162/tacl_a_00619

[14] Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal Recommender Systems: A Survey. *ACM Comput. Surv.* 57, 2, Article 26 (Oct. 2024), 17 pages. doi:10.1145/3695461

[15] Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024. Multimodal Pretraining, Adaptation, and Generation for Recommendation: A Survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona, Spain) *(KDD '24)*. Association for Computing Machinery, New York, NY, USA, 6566–6576. doi:10.1145/3637528.3671473

[16] Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. 2024. Multimodal Pretraining, Adaptation, and Generation for Recommendation: A Survey. arXiv:2404.00621 [cs.IR] https://arxiv.org/abs/2404.00621

[17] Yuqing Liu, Yu Wang, Lichao Sun, and Philip S. Yu. 2024. Rec-GPT4V: Multimodal Recommendation with Large Vision-Language Models. arXiv:2402.08670 [cs.AI] https://arxiv.org/abs/2402.08670

[18] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile) *(SIGIR '15)*. Association for Computing Machinery, New York, NY, USA, 43–52. doi:10.1145/2766462.2767755

[19] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large Language Models are Competitive Near Cold-start Recommenders for Language- and Item-based Preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 890–896. doi:10.1145/3604915.3608845

[20] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. 2024. PMG: Personalized Multimodal Generation with Large Language Models. In *Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) *(WWW '24)*. Association for Computing Machinery, New York, NY, USA, 3833–3843. doi:10.1145/3589334.3645633

[21] Damien Sileo, Wout Vossen, and Robbe Raymaekers. 2022. Zero-Shot Recommendation as Language Modeling. In *Advances in Information Retrieval*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 223–230.

[22] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14918–14937. doi:10.18653/v1/2023.emnlp-main.923

[23] Lei Wang and Ee-Peng Lim. 2024. The Whole is Better than the Sum: Using Aggregated Demonstrations in In-Context Learning for Sequential Recommendation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 876–895. doi:10.18653/v1/2024.findings-naacl.56

[24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang

Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv:2409.12191 [cs.CV] https://arxiv.org/abs/2409.12191

[25] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10052–10065. doi:10.18653/v1/2023.emnlp-main.621

[26] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 3597–3606. doi:10.18653/v1/2020.acl-main.331

[27] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024. A survey on large language models for recommendation. *World Wide Web* 27, 5 (2024), 60.

[28] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. arXiv:2407.10671 [cs.CL] https://arxiv.org/abs/2407.10671

[29] Yuyang Ye, Zhi Zheng, Yishan Shen, Tianshu Wang, Hengruo Zhang, Peijun Zhu, Runlong Yu, Kai Zhang, and Hui Xiong. 2025. Harnessing Multimodal Large Language Models for Multimodal Sequential Recommendation. arXiv:2408.09698 [cs.IR] https://arxiv.org/abs/2408.09698

[30] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (Nov. 2024). doi:10.1093/nsr/nwae403

[31] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL] https://arxiv.org/abs/2306.05685