Retrieval-augmented Recommender System: Enhancing Recommender Systems with Large Language Models

Dario Di Palma
d.dipalma2@phd.poliba.it
Politecnico di Bari
Bari, Italy

ABSTRACT

Recommender Systems (RSs) play a pivotal role in delivering personalized recommendations across various domains, from e-commerce to content streaming platforms. Recent advancements in natural language processing have introduced Large Language Models (LLMs) that exhibit remarkable capabilities in understanding and generating human-like text. RS are renowned for their effectiveness and proficiency within clearly defined domains; nevertheless, they are limited in adaptability and incapable of providing recommendations for unexplored data. Conversely, LLMs exhibit contextual awareness and strong adaptability to unseen data. Combining these technologies creates a powerful tool for delivering contextual and relevant recommendations, even in cold scenarios characterized by high data sparsity. The proposal aims to explore the possibilities of integrating LLMs into RS, introducing a novel approach called Retrieval-augmented Recommender Systems, which combines the strengths of retrieval-based and generation-based models to enhance the ability of RSs to provide relevant suggestions.

KEYWORDS

Large Language Models, Recommender Systems, Prompt Engineering, New Item Recommendation, Hallucination Problem, Retrieval-augmented Recommender System

ACM Reference Format:
Dario Di Palma. 2023. Retrieval-augmented Recommender System: Enhancing Recommender Systems with Large Language Models. In *Seventeenth ACM Conference on Recommender Systems (RecSys '23), September 18–22, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3604915.3608889

1 INTRODUCTION

With the rapid growth of the World Wide Web and vast data, providing users with valuable recommendations became a pivotal challenge. This surge in available information has prompted the development of more complex and accurate models for the recommendation, often leveraging deep learning techniques [7, 14, 20]. From models based on Collaborative Filtering [5, 27], Matrix Factorization [16, 26] and Content-Based [23], the researches on Recommender System (RS) have gradually shifted to Deep Learning-Based RS, leading to modern models relying on Graphs [36, 39], Reinforcement Learning [15, 40] and Transformer-Based techniques [31].

Meanwhile, the introduction of ChatGPT [1] in late November 2022 marks a significant milestone in revolutionizing web interactions for information retrieval, shedding light on the potential of Large Language Models (LLMs). With its robust architecture and extensive training on vast datasets, ChatGPT empowers users to obtain highly relevant and comprehensive answers in moments. As a result, many studies have explored its capabilities [30] and adapted it to specific tasks [4]. Furthermore, there has been a notable surge in investigations to integrate LLMs with RS. For example, Zhiyuli et al. [41] develop BookGPT, a general framework for books recommendation that utilizes LLM. Additionally, Gao et al. [12] introduces Chat-REC, an interactive and explainable Conversational RS that employs LLM. Whereas, Wang et al. [34] put forward the notion of Generative Recommender, employing LLM to comprehend and address users' requirements throughout the recommendation process.

Although ChatGPT demonstrated remarkable capabilities, a thorough investigation is needed to shed light on the true potential of this model. Indeed, Section 4 presents a preliminary exploration of ChatGPT's capabilities in delivering recommendations in terms of accuracy metrics, prompted with the query, "Given a user, as a Recommender System, please provide the top 50 recommendations". The results indicate that ChatGPT (particularly the variant based on GPT-3.5) exhibits comparable performance to state-of-the-art RS models, even without fine-tuning or prompt-engineering techniques. However, additional, comprehensive investigations are needed to establish ChatGPT as a valuable tool supporting the RS community.

This thesis aims to explore multiple research questions in order to examine the potential of Large Language Models (LLMs) in the field of Recommender Systems (RS), with the final goal of introducing a novel perspective on LLMs and ultimately achieving two primary objectives: (i) employing LLMs at the top to create conversational AI integrated with state-of-the-art RSs, and (ii) leveraging LLMs at the lower level to enhance RSs with LLM capabilities. In order to achieve these objectives, the thesis will explore the following research questions:

- (RQ1) Can LLMs provide effective recommendations?
- (RQ2) What are the limitations of using LLMs for recommendations?
- (RQ3) How to handle the recommendation of new items and alleviate the Hallucination Problem of LLMs?

[1] https://openai.com/blog/chatgpt

RecSys '23, September 18–22, 2023, Singapore, Singapore

ACM ISBN 979-8-4007-0241-9/23/09.
https://doi.org/10.1145/3604915.3608889

- (RQ4) Can the combination of retrieval-based (RS) and generation-based (LLM) methods effectively enhance the quality and relevance of the recommendations?

The remainder of this work provides in the next section a concise overview of Large Language Models and prompt engineering. It delves into a comprehensive analysis of the application of LLMs in Recommender Systems. Furthermore, Section 3 covers the research questions and outlines the primary objective of this thesis. Section 4 focuses on preliminary examining ChatGPT's capabilities as a Recommender System. Finally, Section 5 concludes by discussing prospects and aspirations.

2 LARGE LANGUAGE MODELS AND PROMPT ENGINEERING

The introduction of the transformer architecture in "Attention is All You Need" [33] marked a revolutionary advancement in natural language processing. By integrating the self-attention mechanisms, models can capture global dependencies and process text concurrently. Consequently, this breakthrough led to the emergence of notable models such as ELMo [24], GPT [25], and BERT [10], which have made significant contributions to the field. Throughout its development, GPT has undergone extensive training with increasingly large datasets and has benefited from architectural enhancements, leading to the creation of GPT-3. With 175 billion parameters, GPT-3 and other LLMs have demonstrated exceptional capabilities in completing different tasks.

Meanwhile, the research conducted by Brown et al. [6] sheds light on the potential of LLMs in completing tasks through prompt engineering. Indeed, by providing specific instructions, LLMs can assist in accurately answering questions without requiring extensive fine-tuning. In addition, prompt engineering techniques such as few-shot learning [6] and chain-of-thought [35] further amplify the effectiveness of LLMs, enabling significant improvements in complex tasks with minimal effort.

Indeed, through the Reinforcement Learning from Human Feedback (RLHF) [21] and prompt engineering, OpenAI in late November 2022, with the introduction of ChatGPT, a Conversational Agent (CA) initially based on GPT-3, has started a new era for CA. Different companies have started to create and commercialize new powerful LLMs with the aim of developing models that strike a balance between complexity, number of parameters, computational cost, and environmental impact [28].

Despite the impressive performance of models like ChatGPT, LLaMA [32], or PaLM [2] in providing accurate responses across diverse contexts, these models encounter several unresolved challenges. One prominent challenge is their black-box nature, hindering the explanation of their reasoning process. In the realm of recommendations, a major hurdle stems from the extensive training required, such as in the case of ChatGPT, leading to outdated information and limited knowledge about new items beyond 2021.

Nevertheless, the integration of Language Models (LLMs) within recommender systems presents a promising opportunity to address the shortcomings of conventional methodologies. For instance, LLMs can comprehend and analyze natural language, enabling RS to capture user preferences from other perspectives. Indeed, unlike traditional recommender systems that rely heavily on ratings or explicit user feedback, which often suffer from limitations or sparsity, LLMs can overcome these constraints by comprehending implicit user preferences extracted from textual data, such as product reviews, social media posts, or conversational interactions. Combining LLMs with recommender systems can provide more precise and personalized recommendations by harnessing this information, ultimately enhancing the overall user experience.

Indeed, multiple research endeavours have explored the combination of LLMs and recommender systems. For example, recent studies have examined pre-trained language models like BERT [31] or GPT [18] to augment recommendation accuracy. This approach has showcased promising outcomes in diverse domains, including movie recommendations, news article recommendations, and e-commerce product recommendations. The combination of LLMs and recommender systems presents an opportunity to significantly augment the recommendation process, providing users with more satisfying and meaningful experiences and overcoming traditional methodologies' constraints.

3 RESEARCH QUESTIONS AND METHODOLOGY

This thesis aims to explore the potential of foundational models like LLaMA, GPT, and PaLM in the context of Recommendation Systems. Specifically, the aim is to investigate the feasibility of utilizing Large Language Models to develop Retrieval-augmented Recommender Systems. These systems combine retrieval-based and generation-based approaches to provide personalized recommendations to users. Specifically, the research will focus on addressing the following research questions:

(RQ1) Can LLMs provide effective recommendations? To investigate the effectiveness of LLMs in recommendation tasks, initial experiments were performed using ChatGPT. The obtained results are presented in Table 1. To clarify, the experimental setup involved leveraging ChatGPT with the GPT-3.5 model, using the prompt "Given the following user history, give me 50 recommendations" on two datasets, namely MovieLens100K [2] and Facebook Books [3], the results were evaluated based on accuracy.

The findings demonstrate the efficacy of ChatGPT in a Zero-Shot scenario [4] [6], as it successfully generates a curated list of the top 50 ranked items by extracting contextual information from a user's history. The model also seems to capture the similarity between items in the history and offers intriguing recommendations to users. Furthermore, the performance achieved by ChatGPT is comparable with the state-of-the-art models. However, the black-box nature of ChatGPT makes it challenging to understand the specific reasons and mechanisms behind the model's recommendations.

In order to improve the recommendation capabilities of LLMs, the purpose is to explore advanced prompt engineering techniques such as Chain-of-Thought [35] and Tree of Thoughts [37]. These techniques involve carefully designing and refining the prompts given to the model to enhance recommendation accuracy.

[2] https://grouplens.org/datasets/movielens/
[3] https://github.com/sisinflab/LinkedDatasets/tree/master/facebook_book
[4] the zero-shot scenario refers to a situation where a model is able to perform a task or make predictions on classes or categories that it has never seen or been explicitly trained on.

To validate this approach, a comparison will be made between the performance of LLMs with prompt engineering and without prompt engineering, using diverse recommendation datasets. This comparison will help gauge the impact of prompt engineering on the quality and effectiveness of recommendations in light of accuracy.

Additionally, qualitative analyses will be conducted to gain deeper insights into the underlying mechanisms of LLM-based recommendations. This qualitative examination will assess the recommendations' relevance and coherence in specific scenarios. These comprehensive experiments and analyses aim to shed light on the potential of LLMs for recommendation tasks and contribute to a better understanding of their inner workings within recommendation systems.

(RQ2) What are the limitations of using LLMs for recommendations? While prompt engineering techniques can enhance the performance of LLMs by enabling more sophisticated and accurate responses, it is crucial to examine the boundaries of these systems, specifically in the context of recommendations. To address this question, a thorough assessment will examine LLMs' limitations in generating diverse and novel recommendations. Specifically, the focus will be on addressing issues such as popularity bias, which can hinder the exploration of out-of-distribution items, and ensuring serendipity in the recommendation process. Through a systematic analysis of these factors, the objective is to derive valuable insights concerning the practical constraints of employing pure LLMs for recommendations, particularly concerning beyond-accuracy metrics.

(RQ3) How to handle the recommendation of new items and alleviate the Hallucination Problem of LLMs? The recommendation of new items through LLMs represents a significant challenge due to the inherent limitations of their knowledge, which is restricted to the time frame of the training data [38]. Moreover, LLMs are susceptible to the hallucination problem [3], generating fictitious information that lacks accuracy or truthfulness. To address these issues, an alternative approach is required to incorporate new items into the recommendation process without complete retraining of the model. This entails developing effective methods to handle new items and prevent hallucinations.

The primary objective of this investigation is to examine methods for incorporating domain knowledge, as suggested in the works of Emelin et al. [11], Moiseev et al. [19], as well as to explore the application of transfer learning techniques, as presented in the studies Chronopoulou et al. [8], Han et al. [13], alongside finetuning techniques [9]. By harnessing the power of pre-trained models and integrating external knowledge sources, the objective is to enhance the recommendation capabilities of LLMs for unseen and emerging items with the aim of creating a hybrid recommendation model.

Furthermore, to mitigate the risks associated with hallucination [17], strategies will be devised to introduce mechanisms that validate the source of the generated information using information retrieval methods as presented by Shuster et al. [29]. This will involve incorporating fact-checking mechanisms and leveraging user feedback to enhance the reliability of recommendations. By systematically addressing these concerns, the aim is to enhance the robustness and adaptability of LLMs in recommending new items while mitigating the potential risks associated with hallucination.

(RQ4) Can the combination of retrieval-based (RS) and generation-based (LLM) methods effectively enhance the quality and relevance of the recommendations? To answer this question, this research aims to investigate the feasibility of integrating retrieval-based and generation-based approaches in recommender systems using LLMs, leading to a Retrieval-augmented Recommender System. The primary objective is to investigate the feasibility of combining these techniques to enhance the system's overall performance and overcome the commonly encountered cold-start problem [22].

By conducting a systematic analysis, this study will evaluate the advantages and obstacles related to Retrieval-augmented Recommendation Systems (RaRS) employing LLMs. The analysis will entail performing experiments on benchmark datasets and comparing the performance of the proposed approach against existing RS methods. In addition, user studies will be conducted to gather qualitative feedback and assess user satisfaction with the retrieval-augmented recommendation approach. Through investigating these aspects, this research aims to gain a deeper understanding of how LLMs can be effectively employed in recommender systems.

This understanding will facilitate the development of retrieval-augmented recommendation systems that can operate in two ways: (i) utilizing LLMs at the top to create a conversational AI integrated with state-of-the-art RS and (ii) employing LLMs at the bottom to enhance an RS with LLM capabilities.

4 EXPERIMENTAL SETTINGS

To assess the capabilities of the ChatGPT model in addressing recommendation scenarios, an initial experimental analysis was performed to ascertain its potential as a recommender system. By comparing the obtained results with state-of-the-art (SOTA) baselines across diverse application domains, the aim is to provide an initial response to the research question (RQ1) that guides this investigation. The following prompt format was adopted to conduct the experiments: "Given a user, as a Recommender System, please provide the top 50 recommendations. You know that user *[ID]* likes the following *[movies/books]*: *{history of items rated by the user}*." More precisely, employing this prompt for each user in the designated datasets, and taking into account their previous interactions with items, lead ChatGPT to generate a ranked list of the top 50 recommended items, sorted in descending order of relevance. The experiments were carried out using the OpenAI ChatGPT3.5-turbo API [5], to ensure reproducibility the temperature parameter was set to zero, guaranteeing consistent generation of responses and maintaining reproducibility.

To address the token limit imposed by the API, which restricts each message exchanged to 4096 tokens, a prefiltering strategy was employed to handle the user-items interaction through the $k-core$ set to 10 interactions (i.e., all the users and items have at least k recorded interaction). Following this, the datasets were partitioned into separate training and test sets. Specifically, only the training set was utilized to generate prompts for users, whereas the test set was exclusively reserved for metric calculations (i.e., nDCG, HR, and MAP) for three different cutoffs @10, @20, and @50. It is important to note that no additional information was transmitted

[5] https://platform.openai.com/docs/guides/chat

Table 1: Experimental Investigation: Comparing ChatGPT-3.5 with Multiple Baselines on nDCG, HR, and MAP at Cutoff @10, @20, and @50. Best values are in bold. Second best values are underlined. Results are ordered by nDCG@10.

	Model	nDCG@10	HR@10	MAP@10	nDCG@20	HR@20	MAP@20	nDCG@50	HR@50	MAP@50
MovieLens	UserKNN	**0.32358**	0.81711	**0.32792**	**0.31996**	0.87919	**0.27625**	0.35223	0.93624	**0.20770**
	ItemKNN	0.31702	0.81711	0.31869	0.31409	0.87584	0.27227	0.34784	0.94463	0.20649
	RP$^3\beta$	0.31510	**0.82215**	0.31814	0.31754	**0.89094**	0.27204	**0.35393**	**0.94463**	0.20681
	AttributeUserKNN	0.24940	0.73826	0.25819	0.24657	0.80369	0.21759	0.26944	0.87919	0.16451
	EASE^R	0.22447	0.68121	0.23369	0.22404	0.81208	0.19535	0.25142	0.90436	0.14930
	MostPop	0.17172	0.60570	0.17735	0.16748	0.70302	0.14994	0.19229	0.85570	0.11676
	ChatGPT-3.5	0.16927	0.58221	0.18100	0.15198	0.66275	0.14202	0.14081	0.70638	0.09216
	AttributeItemKNN	0.04380	0.29866	0.04426	0.04613	0.40940	0.04064	0.06021	0.61913	0.03609
	VSM	0.02480	0.16107	0.02483	0.02431	0.23993	0.02222	0.03107	0.41443	0.01956
	Random	0.01460	0.11745	0.01468	0.01464	0.18960	0.01361	0.02044	0.34732	0.01260
Facebook Books	**ChatGPT-3.5**	**0.05742**	**0.17267**	**0.02404**	**0.071396**	**0.251286**	**0.020006**	**0.08536**	0.31521	**0.01441**
	AttributeItemKNN	0.05034	0.14254	0.02241	0.05944	0.198384	0.017741	0.06680	0.30860	0.01097
	VSM	0.04592	0.15136	0.01828	0.058051	0.224835	0.016536	0.07117	**0.31668**	0.01243
	AttributeUserKNN	0.04196	0.13446	0.01815	0.050922	0.189566	0.014871	0.06680	0.30860	0.01097
	RP$^3\beta$	0.03097	0.09552	0.01326	0.039529	0.14842	0.010876	0.05010	0.23512	0.00814
	UserKNN	0.03002	0.10213	0.01221	0.040158	0.166054	0.010838	0.05400	0.26231	0.00862
	ItemKNN	0.02873	0.08229	0.01221	0.036906	0.13299	0.009997	0.04890	0.22116	0.00747
	EASE^R	0.02107	0.07568	0.00844	0.031807	0.147686	0.008018	0.05309	0.30639	0.00759
	MostPop	0.00914	0.03233	0.00402	0.011119	0.047024	0.003364	0.03761	0.26672	0.00405
	Random	0.00142	0.00588	0.00053	0.002748	0.016165	0.000622	0.00585	0.04262	0.00083

to ChatGPT, assuming it had acquired sufficient knowledge during its training phase.

The baseline models were trained using a conventional approach, employing training and test datasets within the recommendation framework known as Elliot [1]. Afterwards, a quantitative assessment was performed to gauge the efficacy of recommendations generated by ChatGPT, resulting in the findings presented in Table 1.

To enable ChatGPT to leverage its knowledge for generating recommendations, the prompt was given with the obtained names of each item in the datasets (such as movie titles for MovieLens or book titles for Facebook books). This allowed ChatGPT to utilize its knowledge to give us the recommendations list. Concerning the baselines, the Collaborative Filtering (CF) RSs were provided with information about users' and items' IDs. In contrast, the Content-based Filtering (CBF) RSs learned about the content of items through their genres.

As illustrated in Table 1, by comparing the metrics obtained from ChatGPT and the baselines, it is possible to conclude that ChatGPT can generate recommendations. Moreover, it demonstrates competitive performance with the MovieLens dataset and achieves outstanding results with the Facebook Books dataset. The hypothesis is that ChatGPT has substantial knowledge regarding books, enabling it to function as a powerful recommender system. It can comprehend the context and relationships within users' interaction history to generate relevant recommendations. This experiment validates that ChatGPT is competent in performing recommendation system tasks in a zero-shot scenario.

In evaluating ChatGPT as a Recommender System, a preliminary analysis focused on accuracy metrics reveals the potential of LLMs for Recommendations. In the investigation on Facebook Books, ChatGPT emerged as the front-runner due to its remarkable results in all three metrics. However, when examining the MovieLens dataset, ChatGPT demonstrates comparable performance to other models. Additionally, it is worth noting that the performance of LLMs can be further improved by implementing advanced prompt engineering approaches, as discussed in Section 2. Moreover, it is crucial to recognize the continuous advancement of these systems, as they consistently achieve significant enhancements in their response capabilities over time. These findings show the potential of LLMs and pave the way for new research in the area.

5 CONCLUSION

This thesis aims to investigate the feasibility of integrating retrieval-based and generation-based approaches using Large Language Models (LLMs) to develop a Retrieval-augmented Recommender System (RaRS). The objective is to enhance the overall performance of RSs and overcome the well-known cold-start problem. The first investigation assesses how LLMs can be effectively employed in recommender systems. This exploration has paved the way for the emergence of retrieval-augmented recommendation systems, which can be implemented in two ways: (i) employing LLMs at the top to create conversational AI integrated with state-of-the-art RSs, and (ii) leveraging LLMs at the lower level to enhance RSs with LLM capabilities. Integrating LLMs into RSs offers a powerful means to deliver contextually relevant recommendations, even in scenarios characterized by limited data availability. By combining the strengths of retrieval-based and generation-based models, RSs can overcome limitations and provide personalized suggestions in both familiar and unexplored domains. Continual refinement and enhancement of LLM integration have the potential to revolutionize the field of recommender systems and enhance user experiences across various industries, from e-commerce to content streaming platforms.

Acknowledgements. This work was partially supported by the following projects: Secure Safe Apulia, Casa delle Tecnologie Emergenti Comune di Matera, LUTECH DIGITALE 4.0, OVS Fashion Retail Reloaded, CT_FINCONS_III, KOINÈ, MOST - Centro Nazionale per la Mobilità Sostenibile.

REFERENCES

[1] Vito Walter Anelli, Alejandro Bellogín, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. 2021. Elliot: A Comprehensive and Rigorous Framework for Reproducible Recommender Systems Evaluation. In *SIGIR*. ACM, 2405–2414.

[2] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernández Ábrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, and et al. 2023. PaLM 2 Technical Report. *CoRR* abs/2305.10403 (2023).

[3] Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care* 27, 1 (2023), 1–2.

[4] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. *CoRR* abs/2302.04023 (2023).

[5] John S. Breese, David Heckerman, and Carl Myers Kadie. 2013. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *CoRR* abs/1301.7363 (2013).

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

[7] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah. 2016. Wide & Deep Learning for Recommender Systems. In *DLRS@RecSys*. ACM, 7–10.

[8] Alexandra Chronopoulou, Christos Baziotis, and Alexandros Potamianos. 2019. An Embarrassingly Simple Approach for Transfer Learning from Pretrained Language Models. In *NAACL-HLT (1)*. Association for Computational Linguistics, 2089–2095.

[9] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *CoRR* abs/2305.14314 (2023).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[11] Denis Emelin, Daniele Bonadiman, Sawsan Alqahtani, Yi Zhang, and Saab Mansour. 2022. Injecting Domain Knowledge in Language Models for Task-oriented Dialogue Systems. In *EMNLP*. Association for Computational Linguistics, 11962–11974.

[12] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. 2023. Chat-REC: Towards Interactive and Explainable LLMs-Augmented Recommender System. *CoRR* abs/2303.14524 (2023).

[13] Wenjuan Han, Bo Pang, and Ying Nian Wu. 2021. Robust Transfer Learning with Pretrained Language Models through Adapters. In *ACL/IJCNLP (2)*. Association for Computational Linguistics, 854–861.

[14] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *WWW*. ACM, 173–182.

[15] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement Learning to Rank in E-Commerce Search Engine: Formalization, Analysis, and Application. In *KDD*. ACM, 368–377.

[16] Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *ICDM*. IEEE Computer Society, 263–272.

[17] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38.

[18] Jinming Li, Wentao Zhang, Tian Wang, Guanglei Xiong, Alan Lu, and Gerard Medioni. 2023. GPT4Rec: A Generative Framework for Personalized Recommendation and User Interests Interpretation. *CoRR* abs/2304.03879 (2023).

[19] Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured Knowledge Infusion for Large Language Models. In *NAACL-HLT*. Association for Computational Linguistics, 1581–1588.

[20] Maxim Naumov, Dheevatsa Mudigere, Hao-Jun Michael Shi, Jianyu Huang, Narayanan Sundaraman, Jongsoo Park, Xiaodong Wang, Udit Gupta, Carole-Jean Wu, Alisson G. Azzolini, Dmytro Dzhulgakov, Andrey Mallevich, Ilia Cherniavskii, Yinghai Lu, Raghuraman Krishnamoorthi, Ansha Yu, Volodymyr Kondratenko, Stephanie Pereira, Xianjie Chen, Wenlin Chen, Vijay Rao, Bill Jia, Liang Xiong, and Misha Smelyanskiy. 2019. Deep Learning Recommendation Model for Personalization and Recommendation Systems. *CoRR* abs/1906.00091 (2019).

[21] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

[22] Deepak Kumar Panda and Sanjog Ray. 2022. Approaches and algorithms to mitigate cold start problems in recommender systems: a systematic literature review. *J. Intell. Inf. Syst.* 59, 2 (2022), 341–366.

[23] Michael J. Pazzani and Daniel Billsus. 2007. Content-Based Recommendation Systems. In *The Adaptive Web (Lecture Notes in Computer Science, Vol. 4321)*. Springer, 325–341.

[24] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL-HLT*. Association for Computational Linguistics, 2227–2237.

[25] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[26] Steffen Rendle. 2012. Factorization Machines with libFM. *ACM Trans. Intell. Syst. Technol.* 3, 3 (2012), 57:1–57:22.

[27] Badrul Munir Sarwar, George Karypis, Joseph A. Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *WWW*. ACM, 285–295.

[28] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green AI. *CoRR* abs/1907.10597 (2019).

[29] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. In *EMNLP (Findings)*. Association for Computational Linguistics, 3784–3803.

[30] Chris Stokel-Walker and Richard Van Noorden. 2023. What ChatGPT and generative AI mean for science. *Nature* 614, 7947 (2023), 214–216.

[31] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *CIKM*. ACM, 1441–1450.

[32] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR* abs/2302.13971 (2023).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. 5998–6008.

[34] Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, and Tat-Seng Chua. 2023. Generative Recommendation: Towards Next-generation Recommender Paradigm. *CoRR* abs/2304.03516 (2023).

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.

[36] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-Based Recommendation with Graph Neural Networks. In *AAAI*. AAAI Press, 346–353.

[37] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. *CoRR* abs/2305.10601 (2023).

[38] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know?. In *ACL (Findings)*. Association for Computational Linguistics, 8653–8665.

[39] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. 2018. Graph Convolutional Neural Networks for Web-Scale Recommender Systems. In *KDD*. ACM, 974–983.

[40] Xiangyu Zhao, Long Xia, Liang Zhang, Zhuoye Ding, Dawei Yin, and Jiliang Tang. 2018. Deep reinforcement learning for page-wise recommendations. In *RecSys*. ACM, 95–103.

[41] Aakas Zhiyuli, Yanfang Chen, Xuan Zhang, and Xun Liang. 2023. BookGPT: A General Framework for Book Recommendation Empowered by Large Language Model. *CoRR* abs/2305.15673 (2023).