Leveraging LLMs to Enhance Personality Based Recommender Systems[*]

TAREQ AHRARI, RWTH Aachen University, Germany

Today's world is increasingly dependent on the web and its applications. Recommender systems (RS) have emerged as an important tool for shaping user experience by suggesting relevant items to users across various domains. These systems use techniques like collaborative filtering, content-based filtering, and hybrid filtering. However, these techniques come with challenges, such as the Cold-Start Problem or the Data Sparsity Problem. Personality-based recommender systems (PBRS) address these issues by incorporating personality into recommendations for a personalized user experience, using personality models like the MBTI or Big-Five Model for personality mapping and determining personality through questionnaires. Consequently, efficient personality detection is becoming increasingly important. The rapid development of large language models (LLMs) and their exceptional abilities in solving several tasks implies their use in PBRS by enhancing recommendation and predicting personality traits from text, allowing for the automatization of personality detection. In this paper, we will explore RS, PBRS, and LLMs, evaluate LLMs' ability to detect personality and compare existing approaches to enhance PBRS using LLMs. The goal is to create an overview of the current state of LLM-enhanced PBRS to guide further research with the necessary information.

Additional Key Words and Phrases: Recommender Systems, Personality, Large Language Models, Big-Five, BERT, GPT, recommendation

ACM Reference Format:
Tareq Ahrari. 2024. Leveraging LLMs to Enhance Personality Based Recommender Systems. In *Web Science Seminar SS24.* ACM, New York, NY, USA, Article 1, 15 pages.

1 INTRODUCTION

An integral component of web applications is Recommender Systems (RS) enhancing user experience by providing personalized recommendations. Recommender Systems (RS) are tools to suggest Products and Information to users based on diverse filtering techniques, such as Content-Based Filtering, Collaborative Filtering, and Hybrid Filtering. These Systems are used in various industries, like e-commerce, music, movies, etc., to improve user experience by suggesting items relevant to the user [13, 16]. An example of such a system can be seen in 1. Before we move on to Personality Based Recommender Systems (PBRS), it is important to understand how RS work in general and what issues they face. Formally RS can be described as S the set of all items that can be recommended, U the set of Users receiving these recommendations, F the utility function assesing the value of each items s ∈S for a user u ∈U. F is then $F : U \times S \rightarrow R$ and R is the complete set of recommendations. The goal is to select items for each user to maximize F. Thus, items are assigned rating values, typically integers and a rating matrix of all items can be created. However, since users don't interact with all items, some are not rated at all, hence a filtering technique is needed for either implicit or explicit rating [2, 13].

[*] This document is a student paper produced for a seminar course at RWTH Aachen University. The content, research, and conclusions presented herein are the result of academic exploration by the student author(s) and should be interpreted as such. While this paper uses the ACM template for formatting purposes, it has not been peer-reviewed or published by the Association for Computing Machinery (ACM) or any of its conferences or journals. All errors, omissions, or inaccuracies are the responsibility of the author(s) and any feedback or concerns should be addressed directly to them.

Program: Bachelor/of Science in Computer Science.
Author's Contact Information: Tareq Ahrari, RWTH Aachen University, Aachen, Germany, tareq.ahrari@rwth-aachen.de.

WSS'24, Jul. 2024, RWTH Aachen University
2024.

There are various types of Recommender Systems, but one that stands out especially is personality-based recommender systems (PBRS), leveraging personality traits for more personal and improved recommendations. The recent advancement of Large Language Models (LLM) opens a landscape of possibilities for large data collection and interpretation. As LLMs improve increasingly, more and more human-like language can be generated and understood by LLMs, allowing for a more humanized experience. One such possibility is to enhance PBRS using LLMs, as greater data collection and improved interpretation could be beneficial for personality evaluation. LLMs are already employed to enhance RS, showing promising results [33]. PBRS can generate more personalized recommendations than traditional RS, however, more personal and complex information is needed, leading to ethical concerns like the use of personal data and most importantly, the question of how to map data to personality. Therefore, with the current state of LLMs and the challenges facing PBRS, using LLMs for PBRS could improve recommendations and spark further research, as the topic of PBRS remains relatively unexplored, but LLMs are already engaged in traditional RS [33]. Furthermore, research has already been conducted on personality detection using LLMs [20]. The research goal of this paper can be defined by:

(1) What personality models exist and which one has the highest accuracy for use in recommendation?
(2) How can LLMs detect personality and interpret this data against prominent personality models?
(3) How can LLMs enhance modern PBRS and what challenges do they face?

In light of the increasing importance of good recommendations and the promising results of PBRS and personality detection by LLMs, there is a need to explore the possibility of leveraging LLMs in PBRS. Research has been conducted on this matter, however only on a small scale. Thus, this paper aims to increase awareness of LLM-enhanced PBRS, so that further research can be conducted, as well as provide future researchers with an overview of the current state and approaches, to ease the creation of a taxonomy and guide research on this topic in the right direction.

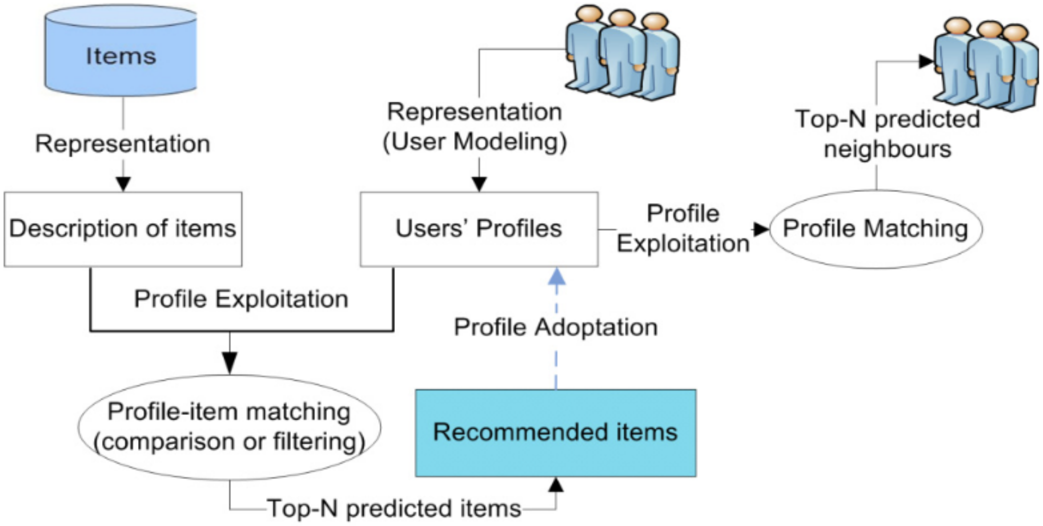


Fig. 1. Framework of the Recommender System Process [13]

2 BACKGROUND INFORMATION

2.1 Personality-based Recommender Systems

Personality-based Recommender Systems (PBRS) are Recommender Systems leveraging Personality traits for improved and more personalized recommendations. For this PBRS work a little bit differently, mainly by having more evaluation phases. First, the user's personality is measured and matched with items based on personality traits, by linking item descriptions with personality types [5]. While RS now use filtering like Collaborative filtering, PBRS computes neighbors and thus person similarity by using personality trait similarity, enhancing the matching and improving performance [5, 17]. PBRS can also deal with issues of traditional RS, like Data Sparsity and the Cold Start Problem, since new items can be rated based on personality features, those unrated items can be assigned values but even new users can be recommended items based on the Personality profile [5, 7]. Since PBRS work with underlying Personality Models, it is important to understand the different models. In the following, we will explain the different models, compare them to each other, and evaluate the best ones. However, the recommendation process also leads to problems. One such Problem is the Cold Start Problem. The cold start Problem is one of the most common problems in RS. The core of it is the addition of a new item or user. So when a new item is introduced, how can it be rated? On the other side, when a new profile is added, how can the preferences of the user be predicted [13]? The second Problem is the Problem of Data Sparsity. If the collection of items available is vast, then the overlap of user ratings or preferences is minimal. Since RS are heavily reliant on user feedback, but many items remain unrated and only a small amount of items are being rated, this poses as a significant problem in predicting preferences for unrated items [2, 16].

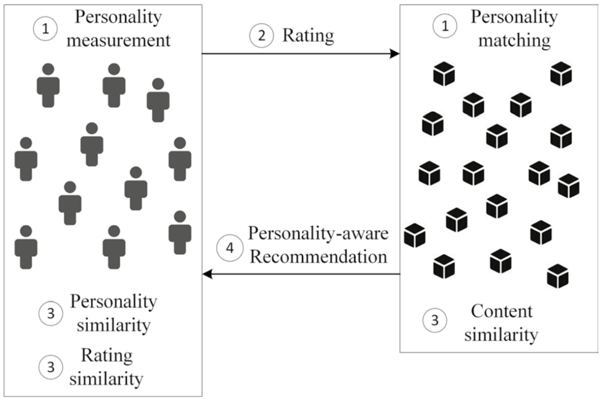


Fig. 2. Personality based Recommender Systems [5]

2.1.1 Personality Models.

Big-Five Model. To measure personality in PBRS, the Big-Five Model is the most used one[5]. The model is based on five factors:

> (I)Surgency (or Extraversion), (II)Agreeableness, (III) Conscientiousness (or Dependability), (IV) Emotional Stability (vs. Neuroticism) and (V) Culture (or Openness) [11].

However, Factor (V) can also be interpreted as Openness or Intellect [11].
In general, there are different Terms for each Dimension but they all still mean the same, it is often dependent on the literature and the preferences of the researcher. To measure personality, PBRS engage every user in a

questionnaire. Different questionnaires are used, like NEO-FFI or NEO-PI-R, but shorter ones like BFI-10 or TIPI are preferred because of their ease of completion. This is also important as users can get bored with too-long questionnaires and make mistakes later on, leading to false results [5].

Item	Question	Dimension
1	I am outgoing, sociable	Extraversion
2	I get nervous easily	Neuroticism
3	I tend to be lazy	Conscientiousness
4	I have an active imagination	Openness
5	I am reserved	Extraversion
6	I am generally trusting	Agreeableness
7	I have few artistic interests	Openness
8	I tend to find fault with others	Agreeableness
9	I do a thorough job	Conscientiousness
10	I am relaxed, handle stress well	Neuroticism

Table 1. An example of a BFI-10 questionnaire [5]

Myers-Briggs Type Indicator. The Myers-Briggs Type Indicator was proposed by Isabel Briggs Myers and Katharine Briggs and uses 4 dimensions, Extraversion-Introversion, Sensing-Intuition, Thinking-Feeling, and Judging-Perceiving. Based on these dimensions, one of each category can be used, and the combination of all 4 creates a type, thus there are 16 types available,
ISFJ, INFP, INFJ, ISTP, ISTJ, ISFP, INTP, INTJ, ENTP, ESFP, ENFP, ESFJ, ESTP, ESTJ, ENFJ and ENTJ [7].

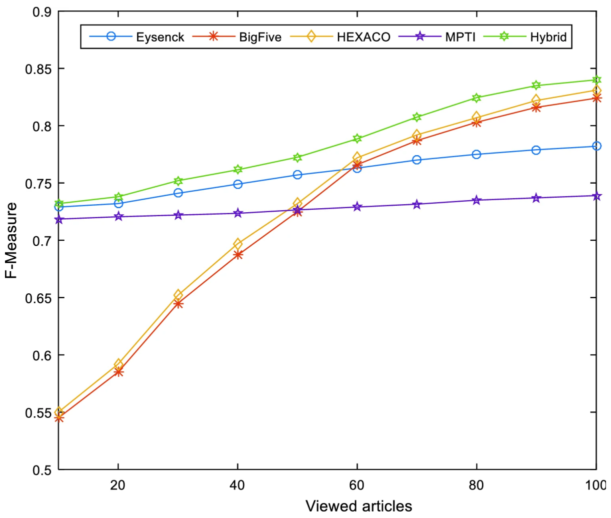


Fig. 3. Comparison of Personality in Recommender Systems [7]

2.2 Large Language Models and Personality Detection

Large language models (LLMs) are mathematical models designed to generate text and understand human language, built on Transformer architecture, and trained on large datasets [21, 32]. In recent years, LLMs have become increasingly popular due to their ability to engage in complex conversations and answer all sorts of questions. LLMs' performance in solving general and complex tasks, and their ability to deal with large datasets, imply their potential for use in RS and PBRS [21, 27]. For this paper, there is no need to know in-depth about LLMs and how they work, but it is important to understand them and their underlying structure. Some notable LLMs addressed in this paper are GPT and BERT. LLMs use a Transformer architecture, and because of its scalability and ability for parallelization, the Transformer architecture by Vaswani et al. has become the most commonly used architecture. Transformers are based on the encoder-decoder structure. The encoder processes the input and transforms it into a continuous representation. The decoder takes the output from the encoder and generates an output sequence [32]. Transformer architecture leverages a mechanism called self-attention, measuring the importance of each token and comparing them to infer contextual information from text [23]. To ensure the order of a sequence, Transformers use positional encoding, which employs sine and cosine functions of different frequencies. One of the first processes is to tokenize the sentence for better parsing in the architecture. LLMs can be pre-trained and fine-tuned, but most commonly prompting and prompting templates are used to guide the generation of answers and thus improve LLMs' performance for specific tasks [32].

As previously stated, LLM's ability to solve general and complex tasks showcases its ability for use in recommendation. However, one important aspect to consider is how LLMs interpret user data using Personality models. This is especially important for PBRS since collected data is always connected to a personality trait, based on the used personality model. Furthermore, personality traits are most often acquired through questionnaires. But it becomes increasingly important to become independent of those because not every person will fill out a questionnaire and even in the case they do, there might be mistakes. To tackle this, approaches could be to interpret preferences, social profiles, or behavior. This kind of information can then be transformed into text and a trained LLM can interpret this data to conclude about personality traits, which is particularly important because it allows for automatization. For that, it is important to understand how exactly LLMs work with user data and use personality models to interpret their findings.

Since the digital Footprint of a person is where a lot of information can be extracted from, mining user interests from social networks could elevate research and possibly allow for automatization. Graphically this can be described as a set of users and a set of topics, there can be a connection between a user and a topic if the user expresses interest in that topic. Now, Topics can be similar to each other, so the user might be interested in those topics as well. Furthermore, some users express the same personality and are modeled as similar users. So if a topic is interesting for one person, it might be interesting for a similar one as well. User personality can be inferred through their online behavior, however, the system also values recent behavior more important than older behavior to increase accuracy. The model uses a Hybrid Filtering Technique to then predict user interest. Based on this approach, the model generates more personalized recommendations. However, personality is not the only measurement for user similarity, the model also measures viewing similarity. The models use Big-Five Personality to assign personality traits to users. Results of the different models and a comparison can be seen in Figure 4 [6].

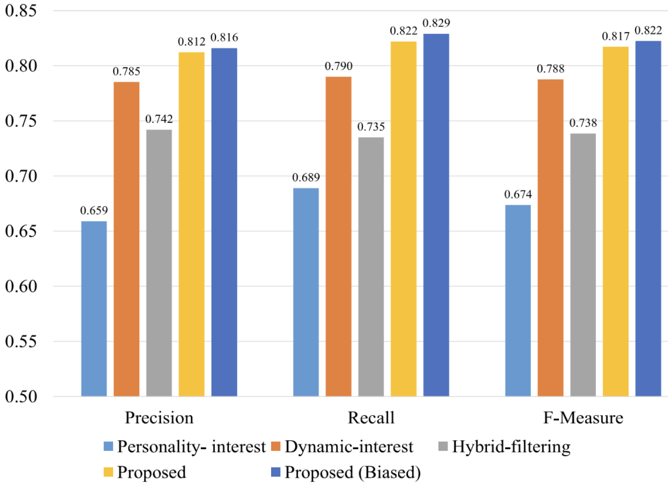


Fig. 4. Results for Personality Extraction [6]

In another approach user's social media profiles are used to infer personality traits based on status updates. In a zero-shot learning environment, LLMs like GPT-3.5 and GPT-4 were used to interpret these status updates and to assign a Big-Five Personality trait. Participants also took a questionnaire that assessed Big-Five personality, the results of the questionnaire and the LLMs were then compared using Pearson correlation coefficient [20]. The results in 5 show a modest correlation for both GPT-3.5 and GPT-4. The highest correlation is for Openness, Extraversion, and Agreeableness, and the lowest for both Neuroticism and Conscientiousness. Overall, GPT-4 also shows increased scores for every trait. The results indicate that LLMs can interpret user data against the Big-Five Theory without being fined-tuned, but systematic biases might arise, highlighting the importance of finding suitable training data. These biases include gender bias, as personality inference is more accurate for women than for men, and also a slight age bias in the GPT-3.5 Model [20].

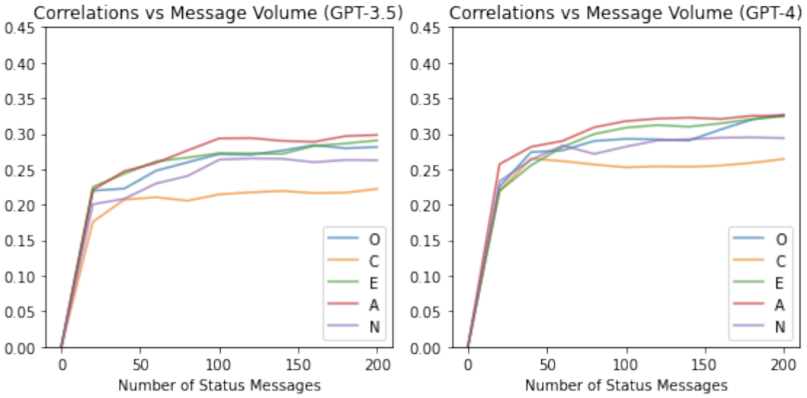
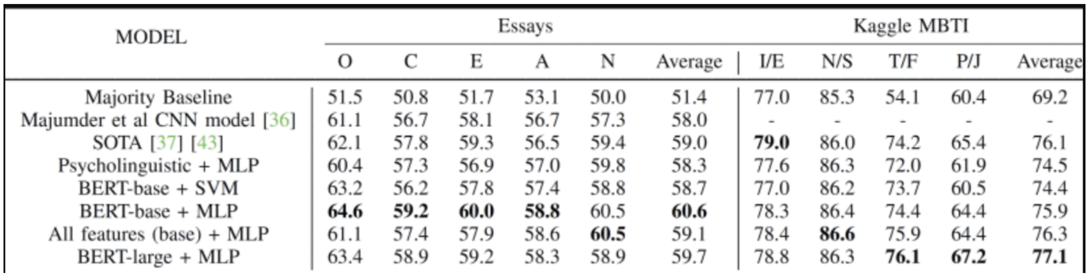


Fig. 5. Big Five in GPT-3.5 and GPT-4 [20]

One other approach is to train a deep learning model using both bottom-up and top-down to predict personality traits from data. This time not only the Big-Five personality traits are used, but also the MBTI traits, using the Essays dataset for Big-Five and the Kaggle dataset for MBTI in text form. The used LLM is BERT. To extract personality, psycholinguistic features, and language model embeddings are used. The psycholinguistic features are derived using the aforementioned text datasets and measured in correlation with personality. Additionally, BERT is used for language model embedding, and although fine-tuned different configurations of BERT are measured and the best one used. Deep Learning Techniques are used to train the model. To analyze the prediction of personality traits based on psycholinguistical features SHAP is used. Although the Model achieves State-of-the-Art performance, Mehta et al. state ethical considerations, as this time personality is not extracted through questionnaires, but through users' online behavior and showcases the potential for extraction without explicit consent. The results can be seen in Figure 6 [18].



MODEL	Essays						Kaggle MBTI				
	O	C	E	A	N	Average	I/E	N/S	T/F	P/J	Average
Majority Baseline	51.5	50.8	51.7	53.1	50.0	51.4	77.0	85.3	54.1	60.4	69.2
Majumder et al CNN model [36]	61.1	56.7	58.1	56.7	57.3	58.0	-	-	-	-	-
SOTA [37] [43]	62.1	57.8	59.3	56.5	59.4	59.0	**79.0**	86.0	74.2	65.4	76.1
Psycholinguistic + MLP	60.4	57.3	56.9	57.0	59.8	58.3	77.6	86.3	72.0	61.9	74.5
BERT-base + SVM	63.2	56.2	57.8	57.4	58.8	58.7	77.0	86.2	73.7	60.5	74.4
BERT-base + MLP	**64.6**	**59.2**	**60.0**	**58.8**	60.5	**60.6**	78.3	86.4	74.4	64.4	75.9
All features (base) + MLP	61.1	57.4	57.9	58.6	**60.5**	59.1	78.4	**86.6**	75.9	64.4	76.3
BERT-large + MLP	63.4	58.9	59.2	58.3	58.9	59.7	78.8	86.3	**76.1**	**67.2**	**77.1**

Fig. 6. Performance on the Essays and Kaggle Dataset [18]

3 RELATED RESEARCH

Although there is still much research to be done on LLM-enhanced PBRS, a vast amount of research for PBRS and Personality Detection using LLMs has been conducted. Research has already shown the potential of PBRS for more personalized recommendations, but also for solving challenges like the Cold Start Problem or Data Sparsity Problem. The first PBRS was developed by Hu et al. and started the research into the area [10]. Since then, PBRS have been developed across various domains, like Movies [25], e-commerce [28] and many more. Asabere et al. proposed a Model called PerSAR, that combines Recommendations with Personality for conference attendees, and used both static and dynamic profiles, as the preferences of attendees can change over time [3]. In a recent approach Wei et al. proposed LLMRec, a graph augmentation recommendation system, that leverages both LLMs to enhance movie recommendation by testing it on Datasets like Netflix or MovieLens [25]. LLMRec employs several strategies for enhanced recommendation, like conducting user node profiling or enhancing item Node Attributes. The Model showcases its superiority over State of the Art Systems with improvements in both Precision and Recall [25]. However, Xu et al. discovered a synergy between the use of LLMs and Machine Learning in Recommendation. Xu et al. use methods like Collaborative Filtering and applies Machine Learning Techniques to it for improvement, while LLMs can be used for "Intelligent Recommendation", like using the user's historical behavior, also allowing for a solution to the Cold Start Problem [28]. One challenge of PBRS is to acquire user personality. The most accurate and most used method is through questionnaires, however, there is a need for a new approach that can extract personality otherwise, like from a user's behavior. This is often the task of LLMs. Tseng et al. used Personas in LLMs and let them Roleplay to different Personas. These Personas were then assessed using Personality Traits. Moreover, the LLMs were adapted to user Preferences to tailor their responses based on the data, and to create a more engaging experience. To predict personality traits, the Big-Five Model was used. The results show that LLMs are effective at Roleplaying Personas and maintain consistency in character portrayal. The study shows the potential for using LLMs for personality detection, as LLMs are engaging well with Personality [22]. Wu (2017) used behavioral features like browsing patterns, content preferences, and social interactions to predict personality traits of the Big-Five Model by applying Machine Learning Techniques. The results show a significant improvement of the traditional RS and an effective way to deal with the Cold Start Problem. Furthermore, extracting a user's implicit behavior yields promising results in both accuracy and user satisfaction [26]. To take it a step further, research was conducted to evaluate personality for cross-domain recommendation. However, with the current state of PBRS, a more modern approach is in using selective ensemble techniques [30]. LLMs can support user data collection, feature engineering, and scoring functions [24]. There are already some LLM-enhanced PBRS, some notable being PALR [29], PAP-REC [14], Health-LLM [12], GIRL [34], LLMRec [25] and Bridging LLMs and Domain-Specific Models for Enhanced Recommendation [24]. PALR uses the users' history to extract behavior patterns and generate recommendations based on this [29]. A challenge of incorporating LLMs in PBRS is prompting, as just slightly different prompts can have significant performance differences. To mitigate immense performance differences, PAP-REC generates personalized automatic prompts using a gradient-based method. However, using this method leads to an extremely large search space, and this to prolonged convergence time [14]. On the other side, Health-LLM showcases the great capability of using large-scale feature extraction to improve accuracy [12]. Another approach, used by GIRL is to use Supervised Fine-Tuning to craft Job Descriptions from CVs and assess the match using Proximal Policy Optimization - based Reinforcement Learning to tailor the output to align with the feedback of the recruiter. Nevertheless, there are still some issues to be dealt with, like community behavior patterns are difficult to express in natural language [34]. Zhang et al. propose an information-sharing module that functions as a connection for collaborative training in LLMs and the domain-specific models [31].

Related research shows the use of LLMs in PBRS, mostly by either using them for recommendations based on prompts or for extracting personality traits.

4 CURRENT STATE OF THE ART PBRS

4.1 Methodology

To measure personality questionnaires are commonly used. These questionnaires will then be compared to assigned personality types in the respective model. Personality similarity can then be computed and similar users can be found, which then get recommended items accordingly. Each PBRS will then be compared by measuring precision, for some also Recall and the F-measure are important. Mathematically Precision can be computed by:

$$P = \frac{TP}{TP + FP} \tag{1}$$

Where TP is the relevant articles viewed by the user, and FP is the irrelevant articles viewed by the user. Furthermore, the Recall can be computed by:

$$R = \frac{TP}{TP + FN} \tag{2}$$

Where FN is the relevant articles not viewed by the user. Precision describes the relevant viewed articles among all articles, and Recall describes the relevant viewed articles among all relevant articles, For a good measurement is important to have both Precision and Recall, thus the focus is on the F-measure, computing the average of both. The F-measure is computed by:

$$F = \frac{2PR}{P + R} \tag{3}$$

In the following sections, the correlation will be measured sometimes, meaning the Pearson correlation coefficient. Moreover, accuracy will sometimes be measured differently, with the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) [7].

4.2 Restaurant recommendation

Christodoulou et al. introduce a PBRS, leveraging user and venue personality, extracted and interpreted by BERT, to improve recommendations for customers and eWOM for topic modeling. The model was fine-tuned using publicly available personality-labeled data. The datasets used are restaurant reviews on TripAdvisor. First, the data from the TripAdvisor Dataset is preprocessed, meaning the removal of unnecessary tokens, like URLs for example. Then the data is tokenized and lowercased. Now, the user-item matrix is created, with the users representing the rows and the columns representing the restaurants. An entry in the matrix is equivalent to the rating of a user for a restaurant. Using eWOM, consumers' opinions are extracted and used for topic modeling. For the modeling, there are both Topics per User and Topics per Restaurant extracted. Topics per Restaurant could be Classy, Clean, Italian, etc., and Topics per User could be Quality of Food, Romantic, Ambiance, etc. At the same time, BERT is used for Personality Classification using two classifiers, one using the Big-Five Model and one for the MBTI Model. They are used to label the personality of the users and the personality of the restaurants. While User's personalities are extracted from their expressions and opinions, restaurant personalities are classified by averaging the personalities of the users who visited them. BERT uses an attention mechanism allowing the Extraction of semantic content by weighing the words [4, 23]. For each dimension of the respective personality model exists a binary classifier. First, the text has to be vectorized, which can be done by using sentence embedding. This vectorized data is then used to fine-tune BERT, which is then ready for use. One Problem for BERT is the processing of texts over 512 tokens since it becomes computationally expensive, so the Model is fine-tuned to deal with long texts. The data used in this study is often longer than 512 tokens, thus naive and semi-naive approaches are employed, because they show the best results. The matrix can now be enhanced by adding personalities and topics to users and restaurants, thus the model is not entirely dependent on personality to make recommendations. Now, there are three matrices, the original one, the one using Big-Five Personalities and User and Restaurant Topics, as well as the one using MBTI Personalities and User and Restaurant Topics. The two personality matrices are trained using an XGBoost Model, which should learn the values of empty entries, therefore predicting a potential rating and thus making recommendations. The two models are compared to the original matrix, and show an improvement in accuracy, with the MBTI Model even outperforming the Big-Five Model [4].

4.3 Fashion recommendation

Mo et al. propose a PBRS for Fashion, using MBTI. First, the SOP Dataset is used and tokenized using BERT. The SOP Dataset consists of Twitter tweets that contain keywords or hashtags of "Personality". Then the model of Mehta et al. is used for personality prediction. The model is pre-trained on the Kaggle Dataset and predicts MBTI personality traits. Liu et al. created a model that can predict a style for an item, it is used on every item in an outfit. The outfit's style is then the style of the item with the most significant score. Similar styles are grouped and styles irrelevant to personality are excluded. Additionally, a Fashion personality is predicted, by first predicting the outfit style and then mapping the ones relevant to the personality of the MBTI. Combining the two predictions and comparing the results with physical compatibility results in a personalized fashion recommendation. The model shows improved accuracy compared to traditional RS and addresses the Cold Start Problem efficiently [15, 18, 19].

4.4 A Chatbot mirroring Personality

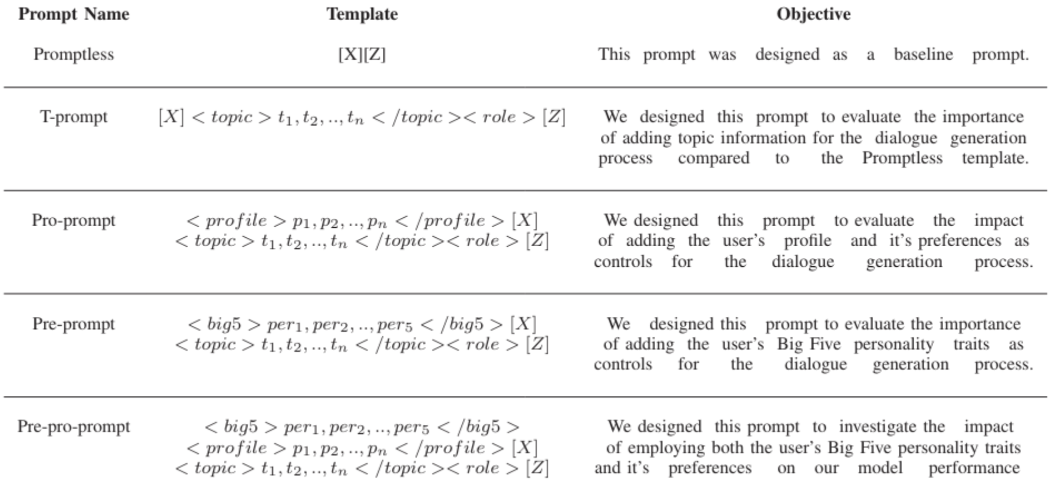
Fernau et al. show another approach to leverage LLMs for PBRS, by training and providing a ChatBot that mimics a user's personality type. Several LLMs are used and compared, like BERT, RoBERTa, etc. The data needs to be cleaned before the model can be trained, for that URLs, newlines, and special characters are removed. Moreover, to increase generalizability, lexical normalization was executed on the model. First, the model is trained using a user's past chat messages. The Model is then fine-tuned using the Personality Cafe database, where the MBTI was already determined. The Chatbot was implemented using Microsoft Azure Bot Framework and was then trained on cues for MBTI types. Since there is no literature about cues and MBTI types, the cues were derived using the Big-Five Model. The Chatbot mimics a Job Recommender Chatbot that interacts with the users. This Chatbot aligning with a user's personality showed high usability and a chance for use in PBRS, as the study also suggests that extracting personality from a text by LLM can be done with only a small effort [9].

4.5 Cross domain PBRS

Using PBRS can not only enhance recommendations but also enable cross-domain recommendations. Using the Big-Five Model, Acharyya and Pervin build a cross-domain PBRS called PEMF-CD, extending the already existing PEMF framework. The model uses Probabilistic matrix factorization to generate the rating matrix and improve it. Then, the matrix of the source domain is factorized, and one of the factorized matrices is a rating matrix, which has missing values and is used to learn these missing values (of the target domain) through optimization. However, two domains can not just be merged like that, otherwise if two non-related domains are merged there might be a negative learning effect. To tackle this challenge, correlation factors can be used to compute the relation between instances of domains and find the ones transfer learning can be applied to. To achieve transfer learning, a Transformer architecture is used. BERT as an LLM is employed for word embedding and personality prediction. First, BERT needs to create an embedded space of texts from both the source and target domain, to create a relationship between both and to enable transfer learning. Then, using the Big-Five Model, personality is predicted. For that, the LLM training itself on the source domain dataset and is then tested on the target domain dataset. After the extraction of personality, PEMF can be used for appropriate recommendations. The Model was pre-trained on the TripAdvisor Dataset and for training and testing the TripAdvisor and Netflix Datasets are used [1].

4.6 Conversational PBRS

Recommendations are not only limited to specific domains but can also apply to conversations. Moreover, LLMs can also be used for the recommendation process instead of personality prediction. For that, Conversational Recommender Systems (CRS) exist, by engaging a user through dialogues. PPPG-DialoGPT is a CRS based on an LLM-enhanced PBRS. This framework combines personality traits and prompt-based learning using DialoGPT, a pre-trained model for dialogues. Prompt-based learning is employed, because LLMs like GPT and BERT are trained on large unstructured Datasets, and using specific templates should improve their performance for specific tasks. Furthermore, personality traits are inferred from user behavior and interactions. DialoGPT is prompted with the individual's personality traits to further enhance recommendation. A template of the prompts can be found in 7. The Model was evaluated using the TG-Redial Dataset for Movie recommendations. The study suggests an improvement over traditional CRS in both efficiency and effectiveness [8].



Prompt Name	Template	Objective
Promptless	[X][Z]	This prompt was designed as a baseline prompt.
T-prompt	$[X] < topic > t_1, t_2, .., t_n < /topic >< role > [Z]$	We designed this prompt to evaluate the importance of adding topic information for the dialogue generation process compared to the Promptless template.
Pro-prompt	$< profile > p_1, p_2, .., p_n < /profile > [X]$ $< topic > t_1, t_2, .., t_n < /topic >< role > [Z]$	We designed this prompt to evaluate the impact of adding the user's profile and it's preferences as controls for the dialogue generation process.
Pre-prompt	$< big5 > per_1, per_2, .., per_5 < /big5 > [X]$ $< topic > t_1, t_2, .., t_n < /topic >< role > [Z]$	We designed this prompt to evaluate the importance of adding the user's Big Five personality traits as controls for the dialogue generation process.
Pre-pro-prompt	$< big5 > per_1, per_2, .., per_5 < /big5 >$ $< profile > p_1, p_2, .., p_n < /profile > [X]$ $< topic > t_1, t_2, .., t_n < /topic >< role > [Z]$	We designed this prompt to investigate the impact of employing both the user's Big Five personality traits and it's preferences on our model performance

Fig. 7. Prompting template for personality traits using Big-Five Model [8]

4.7 Results

For Figure 3 both the Eysenck and HEXACO Model have not been explained, due to these models not taking part in relevant research on this topic. Figure 3 shows two main points. First, in the beginning, both Eysenck and MBTI have a higher value than BigFive and HEXACO. This indicates a better handling of the cold start problem, which is plausible because both use fewer dimensions than BigFive and HEXACO, thus recommendation at the beginning is easier. Eysenck also performs even a little bit better than MBTI, indicated by using even fewer dimensions than MBTI (3 vs 4). However, the more information about a user is available, the better HEXACO and BigFive perform, and even overtake Eysenck and MBTI. While MBTI keeps the same value even with more information, Eysencks value increases a little bit, but after approximately 60 viewed articles, both HEXACO and BigFive overtake Eysenck. Both perform relatively the same, but HEXACO has a slightly higher value. This indicates, that HEXACO and BigFive are better models if more information is present, while Eysenck and MBTI are better with less information and can deal with the Cold Start problem. The graph also shows a hybrid system, which at all stages performs best, showcasing the strength of a hybrid system. The hybrid model is a mix of the Big-Five, Eysenck, HEXACO, and MBTI Model and leverages all their advantages to create a Model that can both effectively handle the Cold-Start Problem and the Data Sparsity Problem [7].

The graph in figure 4 shows that just using personality interest performs weakly in comparison to dynamic and hybrid-filtering models. Moreover, using a model that incorporates both user personality and viewing similarity yields the highest results for accuracy, showcasing that personality alone is not good enough to compute similar users. Moreover, enhancing the Model using an LLM yields the highest result, exhibiting the strength of the use of LLMs in PBRS. A Model that also contains bias performs even slightly better. The bias Model would take the user's dominant personality type and associate characters related to that trait. These characters can be associated with items. For example, a Person expressing high Openness would be classified as Artistic, and Articles related to Art can be recommended to such users. Articles can be classified by keywords, labels, and analyzing the main text [6].

The results in 5 indicate that LLMs can interpret user data against the Big-Five Theory without being fined-tuned, however, the correlation although positive doesn't reach statistical significance. Furthermore, in this experiment, an unsupervised model was used, with supervised models reaching higher correlation, although still not one of statistical significance [20].

The results in 6 show stronger performance in comparison to traditional psycholinguistic features. This indicates that next to GPT also BERT can improve personality trait extraction, making its use particularly interesting in PBRS [18].

The results in table 2 show the different LLM enhanced PBRS, their used personality model, and the achieved accuracy. The Cross-domain PBRS didn't exactly compute an accuracy value but measured MAE and RMSE to infer accuracy. As the MAE and RMSE values are relatively low, especially in comparison to other approaches for CRS, an improved accuracy is evident [1]. All LLM-enhanced PBRS achieved an accuracy of significance, and all performed better than respective traditional RS [4, 8, 9, 19]. Moreover, it is apparent that for different domains there is a significant difference in accuracy. Using the Big-Five Model, CRS recommendation performs significantly better than Restaurant recommendation (a difference of 0,206 or roughly 20%). Furthermore, for the MBTI Model accuracy is not as high as for the Big-Five Model, but a significant variance is still apparent, while the Restaurant recommendation publishes a high accuracy, the Fashion recommendation performs slightly worse, and the Chatbot even worse (the highest difference is 0,229 or roughly 23%). The CRS recommender performs best at all stages, showcasing the strength to use LLMs for recommendation incorporating both personality traits as well as doing the recommendation itself, but also highlighting the importance of accurate prompting.

Framework	Personality Model	Accuracy	MAE	RMSE
Restaurant recommendation [4]	MBTI	0.839	/	/
Restaurant recommendation [4]	Big-Five	0.698	/	/
Fashion recommendation [19]	MBTI	0.7676	/	/
Chatbot [9]	I/E (MBTI)	61% (0.61)	/	/
Cross domain recommendation [1] (80:20 train-test split)	Big-Five	/	1.208	6.540
CRS recommendation [8]	Big-Five	0.9040	/	/

Table 2. An overview of LLM enhanced PBRS

Research shows the ability to incorporate personality in RS, dealing effectively with open challenges and allowing for a more personalized experience [3]. Furthermore, research shows the increasing improvement of LLM for accurate personality detection from text. This can prove especially useful since personality detection is commonly done by using questionnaires, which however are prone to mistakes and require an active engagement by the user. By extracting personality traits from text, LLMs show potential for automatization of this very problem, as previous conversations, social media posts, and online behavior can be transformed into text and a trained LLM can then predict personality. However, by using a graph of users and topics the results in Figure 4 show that it is also important to incorporate viewing similarity, as similar personality alone is not enough for similar interests in users. Since enhancing PBRS with LLMs is a relatively new field, there hasn't been too much research yet, however existing Literature already shows that when LLMs are used they can enhance recommendation, for example by personality detection, as it allows for separation of this task, and the recommendation engine can still focus on recommending rather than on detecting personality, also allowing for higher usability in different domains. Other approaches are to let the LLM recommend or mimic a person's personality [6, 18, 20]. For LLM-enhanced

PBRS it is apparent that the use of different personality models yields different results. This aligns with the findings of 3, where the two in this paper measured personality models, Big-Five and MBTI, when employed in the same framework have different results. Moreover, a higher accuracy is achieved by the Big-Five Model, which is also apparent in 3. These findings should indicate that the personality model is quite important when used, as a significant difference can occur [7]. Additionally, for every domain, different models perform better. As a result, the field of LLM-enhanced PBRS still needs a lot of attention from researchers, but the potential is already clear, as digital Footprints increase (and with that, also the available data on a user), as well as the rapid evolution of LLMs is happening, enabling every new model with possibly better abilities.

4.8 Research Questions

After analyzing our results we can now answer our research questions.
For the first question, our results showcase that the HEXACO and Big-Five Models perform best when there is more information available, while the Eysenck and MBTI Models perform better when there is less information available. This showcases the strength of both MBTI and Eysenck to deal with the Cold Start Problem, while the HEXACO and Big-Five Models deal better with the Data Sparsity Problem. Moreover, a hybrid approach performs even better than those two, tackling both the Cold Start Problem and Data Sparsity Problem. However, the Bulk of research focuses on the Big-Five and MBTI Models, while other Models remain unexplored [7].

LLMs detect personality based on user profiles, behavior, and conversations. This is then turned into text, and specific prompts are used to detect personality. Using an underlying Model expressions and behavior of a user can be assigned to a dimension of a personality model. In the end, a personality classification of the user exists, whereas, in the case of MBTI, a clear assignment to a type is made. In the case of the Big-Five Model, every dimension is assigned values, like high or low for example. Furthermore, research shows that both the MBTI and Big-Five Models are the most used because MBTI allows for a binary classification and Big-Five allows for an easy categorization [6, 7, 18, 20].

To answer the last question, we investigated recent LLM-enhanced PBRS and focused on those who use LLMs for personality extraction rather than for recommendation alone. When LLMs are employed for the extraction and detection of personality, they can effectively deal with the Data Sparsity problem because they can make use of large amounts of data. Moreover, LLMs can also deal with the Cold Start Problem as personality is measured before the first recommendation. The results indicate the strength of combining LLMs and PBRS for Personality Detection. This would also solve a big problem in current PBRS, as personality is often measured through questionnaires, which can have mistakes and users might be hesitant to fill out a questionnaire for a website. Hence, new approaches to measuring personality are needed, and LLMs can provide a solution to this problem, by extracting personality from a user's digital footprint, and then using this data for recommendations [1, 4, 8, 9, 19].

5 OPEN RESEARCH OPPORTUNITIES AND FUTURE RESEARCH

Research on LLM-enhanced PBRS has yet to be conducted in a wider range and has to continuously be updated as LLM develops fast and every new model is an improvement of the old one. Furthermore, a vast amount of LLMs exist and many models haven't had the chance yet to prove themselves in PBRS. The same also applies to Personality Theories, as the bulk of the research focuses on the Big Five and MBTI models, but other Theories and Hybrid approaches remain unexplored. Evaluation of LLM-enhanced PBRS often resolves around recommendation accuracy and their dealing with common challenges like the Cold Start Problem or the Data Sparsity Problem. However, many more aspects have to be considered and evaluated, like user acceptance, security, and data privacy. Furthermore, models are rarely tested on several datasets, and Biases are often disregarded, resulting in one-dimensional results. One other important aspect is the cost of using LLMs, as research only tests the accuracy of these Models, but does not measure the Costs these Models could have. While some Models take into consideration that preferences can change over time, Personality change was never measured, however, this is important to understand a person's changing preferences. New methods have to deal with that. Moreover, ethical concerns are rarely addressed, although these systems deal with sensitive data. Much research has to be done before this kind of PBRS should enter sophisticated markets or environments, to fully understand their limitations and to apply them correctly.

Further research needs to be done on this topic. Some questions that should be picked up by future research are:

(1) How does a hybrid personality model for a LLM enhanced PBRS perform?
(2) What are the costs of using LLMs in PBRS?
(3) Can Biases be mitigated in LLM-enhanced PBRS?
(4) Can LLM-enhanced PBRS detect changing Personalities over time?
(5) In which domains do the different personality models perform best?

6 CONCLUSION

This paper investigates the current state of LLM-enhanced PBRS and presents an overview of state-of-the-art approaches. Our results show that LLMs can predict personality from text when trained and fine-tuned correctly with a significant correlation. Incorporating LLMs into PBRS, specifically for personality detection, allows for a possible automatization by leveraging a user's digital footprint, and shows promising results for enhanced and more personalized recommendations.

REFERENCES

[1] Somdeep Acharyya and Nargis Pervin. 2023. *Enhancing Cross-Domain Recommendations: Leveraging Personality-Based Transfer Learning with Probabilistic Matrix Factorization.* https://doi.org/10.2139/ssrn.4565894

[2] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[3] Nana Yaw Asabere, Amevi Acakpovi, and Mathias Bennet Michael. 2018. Improving Socially-Aware Recommendation Accuracy Through Personality. *IEEE Transactions on Affective Computing* 9, 3 (2018), 351–361. https://doi.org/10.1109/TAFFC.2017.2695605

[4] Evripides Christodoulou, Andreas Gregoriades, Herodotos Herodotou, and Maria Pampaka. [n. d.]. Combination of user and venue personality with topic modelling in restaurant recommender systems. *16130073* 3219 ([n. d.]), 21–36. https://ktisis.cut.ac.cy/handle/20.500.14279/29890

[5] Sahraoui Dhelim, Nyothiri Aung, Mohammed Amine Bouras, Huansheng Ning, and Erik Cambria. 2022. A survey on personality-aware recommendation systems. *Artificial Intelligence Review* 55, 3 (2022), 2409–2454. https://doi.org/10.1007/s10462-021-10063-7

[6] Sahraoui Dhelim, Nyothiri Aung, and Huansheng Ning. 2020. Mining user interest based on personality-aware hybrid filtering in social networks. *Knowledge-Based Systems* 206 (2020), 106227. https://doi.org/10.1016/j.knosys.2020.106227

[7] Sahraoui Dhelim, Liming Chen, Nyothiri Aung, Wenyin Zhang, and Huansheng Ning. 2023. A hybrid personality-aware recommendation system based on personality traits and types models. *Journal of Ambient Intelligence and Humanized Computing* 14, 9 (2023), 12775–12788. https://doi.org/10.1007/s12652-022-04200-5

[8] Fahed Elourajini and Esma Aïrncur. 2023. PPPG-DialoGPT: A Prompt-based and Personality-aware Framework For Conversational Recommendation Systems. In *2023 IEEE International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. IEEE, 273–279. https://doi.org/10.1109/WI-IAT59888.2023.00044

[9] Daniel Fernau, Stefan Hillmann, Nils Feldhus, and Tim Polzehl. 2022. Towards Automated Dialog Personalization using MBTI Personality Indicators. In *Interspeech 2022*. ISCA, ISCA. https://doi.org/10.21437/interspeech.2022-376

[10] Rong Hu and Pearl Pu. 2011. Enhancing collaborative filtering systems with personality information. In *Proceedings of the fifth ACM conference on Recommender systems - RecSys '11*, Bamshad Mobasher, Robin Burke, Dietmar Jannach, and Gediminas Adomavicius (Eds.). ACM Press, New York, New York, USA, 197–204. https://doi.org/10.1145/2043932.2043969

[11] Steven Hyman (Ed.). 2002. *Personality and Personality Disorders: The Science of Mental Health* (first edition ed.). Routledge, Boca Raton, FL. https://permalink.obvsg.at/

[12] Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Du Mengnan, and Yongfeng Zhang. [n. d.]. Health-LLM: Personalized Retrieval-Augmented Disease Prediction System. http://arxiv.org/pdf/2402.00746v6

[13] Shah Khusro, Zafar Ali, and Irfan Ullah. 2016. Recommender Systems: Issues, Challenges, and Research Opportunities. In *Information Science and Applications (ICISA) 2016*, Kuinam J. Kim and Nikolai Joukov (Eds.). Lecture notes in electrical engineering, Vol. 376. Springer Singapore, Singapore, 1179–1189. https://doi.org/10.1007/978-981-10-0557-2{_}112

[14] Zelong Li, Jianchao Ji, Yingqiang Ge, Wenyue Hua, and Yongfeng Zhang. [n. d.]. PAP-REC: Personalized Automatic Prompt for Recommendation Language Model. http://arxiv.org/pdf/2402.00284v1

[15] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 27.06.2016 - 30.06.2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1096–1104. https://doi.org/10.1109/CVPR.2016.124

[16] Linyuan Lü, Matúš Medo, Chi Ho Yeung, Yi-Cheng Zhang, Zi-Ke Zhang, and Tao Zhou. 2012. Recommender systems. *Physics Reports* 519, 1 (2012), 1–49. https://doi.org/10.1016/j.physrep.2012.02.006

[17] Xinyuan Lu and Min-Yen Kan. [n. d.]. Improving Recommendation Systems with User Personality Inferred from Product Reviews. http://arxiv.org/pdf/2303.05039v2

[18] Yash Mehta, Samin Fatehi, Amirmohammad Kazameini, Clemens Stachl, Erik Cambria, and Sauleh Eetemadi. 11/17/2020 - 11/20/2020. Bottom-Up and Top-Down: Predicting Personality with Psycholinguistic and Language Model Features. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1184–1189. https://doi.org/10.1109/ICDM50108.2020.00146

[19] Dongmei Mo, Xingxing Zou, and Wai Keung Wong. 2023. Supplementary Material: Personalized Fashion Recommendation via Deep Personality Learning. (2023).

[20] Heinrich Peters and Sandra Matz. [n. d.]. Large Language Models Can Infer Psychological Dispositions of Social Media Users. http://arxiv.org/pdf/2309.08631v1

[21] Murray Shanahan. 2024. Talking about Large Language Models. *Communications of the ACM* 67, 2 (2024), 68–79. https://doi.org/10.1145/3624724

[22] Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Yu-Ching Hsu, Jia-Yin Foo, Chao-Wei Huang, and Yun-Nung Chen. [n. d.]. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. http://arxiv.org/pdf/2406.01171v1

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. [n. d.]. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

[24] Arpita Vats, Vinija Jain, Rahul Raja, and Aman Chadha. [n. d.]. Exploring the Impact of Large Language Models on Recommender Systems: An Extensive Review. http://arxiv.org/pdf/2402.18590v3

[25] Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. 2024. LLMRec: Large Language Models with Graph Augmentation for Recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, Luz Angélica, Silvio Lattanzi, Andrés Muñoz Medina, Leman Akoglu, Aristides Gionis, and Sergei Vassilvitskii (Eds.). ACM, New York, NY, USA, 806–815. https://doi.org/10.1145/3616855.3635853

[26] Wen Wu. 2017. Implicit Acquisition of User Personality for Augmenting Recommender Systems. In *Companion Proceedings of the 22nd International Conference on Intelligent User Interfaces*, George A. Papadopoulos, Tsvi Kuflik, Fang Chen, Carlos Duarte, and Wai-Tat Fu (Eds.). ACM, New York, NY, USA, 201–204. https://doi.org/10.1145/3030024.3038287

[27] Lanling Xu, Junjie Zhang, Bingqian Li, Jinpeng Wang, Mingchen Cai, Wayne Xin Zhao, and Ji-Rong Wen. [n. d.]. Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis. http://arxiv.org/pdf/2401.04997v1

[28] Xiaonan Xu, Yichao Wu, Penghao Liang, Yuhang He, and Han Wang. [n. d.]. Emerging Synergies Between Large Language Models and Machine Learning in Ecommerce Recommendations. http://arxiv.org/pdf/2403.02760v2

[29] Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. [n. d.]. PALR: Personalization Aware LLMs for Recommendation. http://arxiv.org/pdf/2305.07622v3

[30] Xu Yu, Qinglong Peng, Lingwei Xu, Feng Jiang, Junwei Du, and Dunwei Gong. 2021. A selective ensemble learning based two-sided cross-domain collaborative filtering algorithm. *Information Processing & Management* 58, 6 (2021), 102691. https://doi.org/10.1016/j.ipm.2021.102691

[31] Wenxuan Zhang, Hongzhi Liu, Du Yingpeng, Chen Zhu, Yang Song, Hengshu Zhu, and Zhonghai Wu. [n. d.]. Bridging the Information Gap Between Domain-Specific Model and General LLM for Personalized Recommendation. http://arxiv.org/pdf/2311.03778v1

[32] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Du Yifan, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. [n. d.]. A Survey of Large Language Models. http://arxiv.org/pdf/2303.18223v13

[33] Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. [n. d.]. Recommender Systems in the Era of Large Language Models (LLMs). http://arxiv.org/pdf/2307.02046v5

[34] Zhi Zheng, Zhaopeng Qiu, Xiao Hu, Likang Wu, Hengshu Zhu, and Hui Xiong. [n. d.]. Generative Job Recommendations with Large Language Model. http://arxiv.org/pdf/2307.02157v1