software

MDPI

Article

Precision-Driven Product Recommendation Software: Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems

Konstantinos I. Roumeliotis [1,*], Nikolaos D. Tselikas [1] and Dimitrios K. Nasiopoulos [2]

[1] Department of Informatics and Telecommunications, University of Peloponnese, Akadimaikou G. K. Vla-chou Street, 22131 Tripoli, Greece; ntsel@uop.gr

[2] Department of Agribusiness and Supply Chain Management, School of Applied Economics and Social Sciences, Agricultural University of Athens, 11855 Athens, Greece; dimnas@aua.gr

[*] Correspondence: k.roumeliotis@uop.gr; Tel.: +30-271-037-2216

Abstract: This paper presents a pioneering methodology for refining product recommender systems, introducing a synergistic integration of unsupervised models—K-means clustering, content-based filtering (CBF), and hierarchical clustering—with the cutting-edge GPT-4 large language model (LLM). Its innovation lies in utilizing GPT-4 for model evaluation, harnessing its advanced natural language understanding capabilities to enhance the precision and relevance of product recommendations. A flask-based API simplifies its implementation for e-commerce owners, allowing for the seamless training and evaluation of the models using CSV-formatted product data. The unique aspect of this approach lies in its ability to empower e-commerce with sophisticated unsupervised recommender system algorithms, while the GPT model significantly contributes to refining the semantic context of product features, resulting in a more personalized and effective product recommendation system. The experimental results underscore the superiority of this integrated framework, marking a significant advancement in the field of recommender systems and providing businesses with an efficient and scalable solution to optimize their product recommendations.

Keywords: recommender systems; recommender system algorithms; product recommendation; product recommendation algorithms; GPT model; k-means clustering; content-based filtering; hierarchical clustering; recommender systems evaluation; model evaluation

check for updates

Citation: Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Precision-Driven Product Recommendation Software: Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems. *Software* **2024**, *3*, 62–80. https://doi.org/10.3390/software3010004

Academic Editor: Paolino Di Felice

Received: 4 February 2024
Revised: 22 February 2024
Accepted: 27 February 2024
Published: 29 February 2024

1. Introduction

The cart phase represents a pivotal juncture in e-commerce, serving as both the moment of truth and an opportune moment for upselling. Personalized product recommendations during this stage are crucial for elevating the average order value (AOV) without compromising conversion rates [1].

Evaluating recommender systems has consistently posed a challenge. Several metrics, including user surveys, accuracy metrics, user engagement metrics, precision and recall, and A/B testing, have been introduced as methods to assess recommendation systems. While several evaluation metrics are available, the subjective judgment of customers emerges as the crucial factor in determining the authentic success of a recommender system. Their discernment regarding the value of a product recommendation becomes the ultimate benchmark for the system's effectiveness. As such, striking a balance between quantitative metrics and customer satisfaction remains a paramount consideration in the ongoing refinement of recommender systems.

In this study, we propose an innovative methodology to elevate precision within product recommendation systems through the integration of advanced unsupervised machine learning models with the state-of-the-art GPT-4 large language model (LLM). Our investigation focuses on three unsupervised models—specifically, K-means clustering, content-based filtering (CBF), and hierarchical clustering—all meticulously trained to enhance the accuracy of product recommendations. To evaluate the effectiveness and conduct comprehensive comparisons among these models, we harness the robust capabilities of the GPT-4 model [2].

The devised methodology incorporates the development of a user-friendly flask-based API, strategically designed to provide e-commerce owners with a seamless solution for the integration and optimization of their product recommendation systems. Leveraging this API, e-commerce proprietors can effortlessly input their product data into CSV format, initiating an automated process that encompasses the training and evaluation of the three unsupervised models. The pivotal inclusion of the GPT-4 LLM introduces a groundbreaking dimension to our approach, enabling a sophisticated understanding and refinement of the semantic context associated with product features. This augmentation significantly amplifies the precision and relevance of the recommendations provided.

In practical applications, our proposed system facilitates user interaction by allowing e-commerce owners to input a product title via the API. Subsequently, the software employs the trained models and GPT-4 LLM to deliver the most contextually fitting product recommendations. This pioneering framework not only streamlines the implementation of intricate unsupervised models but also capitalizes on the advanced natural language understanding capabilities inherent to GPT-4, resulting in refined and personalized product recommendations.

The empirical results derived from our extensive experiments underscore the superiority of employing a language model (LLM) as an evaluation tool over the time-consuming and cost-inefficient process of human evaluation. Our proposal involves utilizing the GPT-4 model as an evaluation tool to refine the precision of recommendations generated by traditional unsupervised algorithms. This is achieved through multiple rounds of training and evaluations on these models, with adjustments made to model parameters or hyperparameter tuning, all aimed at enhancing recommendation precision.

This paper significantly contributes to the ever-evolving landscape of recommender systems by introducing an efficient and scalable solution. It empowers e-commerce enterprises to optimize their product recommendation systems by seamlessly integrating state-of-the-art machine learning and natural language processing technologies.

The primary aim of this research is multifold: firstly, to assess and compare the performance of the supervised models for recommendation tasks; secondly, to introduce the GPT-4 LLM for model evaluation; and thirdly, to address specific research questions that have not been adequately answered by prior studies:

Q1: Which unsupervised recommender system algorithm demonstrates superior efficacy in product recommendation tasks?

Q2: Is the GPT-4 model capable of evaluating the predictions of traditional unsupervised models?

Q3: Can LLMs replace human evaluations?

Q4: Do NLPs have the capability to evaluate recommendations from unsupervised learning algorithms?

Q5: Can the proposed flask-based API contribute to the accessibility and usability of sophisticated recommendation systems for e-commerce owners?

To address the aforementioned research questions, the paper begins by providing a concise literature review in Section 2. Section 3 outlines the research and development methodology, encompassing the training of unsupervised models and their evaluation using the GPT model. Section 4 delves into the results, extracting insights and formulating statements that address the research questions. Finally, in Section 5, an alternative traditional method is presented for assessing the effectiveness of recommender systems, drawing comparisons with our proposed GPT-based evaluation tool.

2. Literature Review

In the ever-evolving realm of e-commerce, the curation and presentation of recommended products have traditionally rested on the shoulders of e-commerce owners, entailing a manual and intricate process across various online platforms, encompassing both product pages and shopping carts. E-commerce proprietors historically undertook the laborious task of manually selecting products believed to align with the current visitor's preferences or proposing complementary items strategically timed just before the checkout stage [3]. This manual curation spanned a spectrum of strategies, including cross-selling, up-selling, bundle recommendations, frequently-bought-together suggestions, and personalized recommendations. The associated workload for e-commerce owners in executing these strategies was undeniably colossal.

However, the landscape of e-commerce underwent a paradigm shift with the advent of machine learning and artificial intelligence, ushering in a transformative era. This era introduced automated solutions through models and algorithms meticulously trained for the explicit purpose of handling the intricacies of recommendation systems [4]. Initially adopted by major marketplaces, these models have since evolved and adapted to augment and streamline the product recommendation processes within individual e-commerce platforms. Efficient e-commerce performance is also crucial for real-time processing and responsiveness, which are essential factors in systems that provide recommendations, especially in dynamic and data-intensive environments [5,6].

This Section 2 embarks on an exploration of the profound evolution brought about by machine learning and AI in the realm of e-commerce product recommendation systems. It delves into the most renowned product recommendation strategies, elucidating how these strategies have been transformed and optimized through the integration of automated technologies. Moreover, the review aims to present the most effective algorithms tailored for each specific recommendation strategy, shedding light on the efficiency, accuracy, and adaptability of these intelligent systems in catering to the diverse needs of online shoppers.

2.1. Personalized Product Recommendation Strategies

In the realm of e-commerce, strategic approaches such as cross-selling play a pivotal role in maximizing customer value and satisfaction. Specifically, complementary product recommendations form a key component of this strategy, where businesses suggest products that complement those already in the customer's shopping cart [7]. For example, if a customer adds a camera to their cart, cross-selling might involve recommending accessories like lenses or tripods. This not only enhances the overall shopping experience by providing relevant options but also opens avenues for the seller to increase the average transaction value and build a more comprehensive relationship with the consumer [8]. The thoughtful execution of a cross-selling recommender system not only boosts sales but also fosters customer loyalty by showcasing a genuine understanding of their needs [9].

Conversely, up-selling is another strategic sales tactic in e-commerce that focuses on encouraging customers to consider higher-value or superior-quality alternatives to the items in their cart [10]. By suggesting premium product alternatives, such as an advanced smartphone model with enhanced features, businesses aim to entice customers to spend more [11]. This approach not only contributes to increased revenue but also elevates customer satisfaction and loyalty by highlighting the added value associated with premium offerings [9]. Through an effective up-selling recommender system, customers are provided with an opportunity to explore superior options and enhance their overall shopping experience.

Moving on to bundle recommendations, this approach involves suggesting sets of products that include items already in the customer's cart along with additional related products [12]. For instance, if a customer selects a camera and accessories, a bundle recommendation might offer a complete photography kit with added items like a camera bag and cleaning kit. This strategy simplifies the shopping process for customers, offering a ready-made solution and often presenting a cost-saving opportunity compared to individual purchases [13]. Bundle recommender systems enhance the overall shopping experience by catering to the customer's needs in a holistic manner, contributing to increased sales and customer satisfaction [14].

The frequently-bought-together strategy, rooted in collaborative filtering algorithms and historical purchasing patterns, suggests products commonly bought in tandem with the items in the customer's cart [15]. By analyzing previous customer preferences using user profiling techniques, this strategy offers companion products or popular pairings. For example, if a customer selects a laptop, the system might recommend commonly paired accessories like a laptop bag or wireless mouse. This data-driven approach streamlines decision making, enhances the shopping experience, and increases the likelihood of up-sells and additional purchases, contributing to a more satisfying overall experience [16].

In the era of modern e-commerce, personalized recommendations stand as a cornerstone, leveraging customer data, preferences, and behavior to offer tailored and customized product suggestions [13]. This sophisticated approach analyzes diverse factors such as past purchases, browsing history, and demographic information to create a highly personalized shopping experience. By providing unique recommendations aligned with each customer's specific interests and needs, businesses aim to enhance customer engagement and satisfaction [17]. Whether suggesting similar products based on past purchases or offering personalized promotions, this strategy not only contributes to a more enjoyable and efficient shopping experience but also fosters customer loyalty through a thoughtful and individualized approach [9].

2.2. Product Recommendation Models and Algorithms

In the ever-evolving realm of e-commerce, the integration of machine learning (ML) and artificial intelligence (AI) models, particularly in recommender systems, is ubiquitous. These models play a crucial role in providing a wide array of dynamic and efficient product recommendations. Each recommendation strategy exhibits distinctive characteristics, utilizing varying datasets, features, and objectives. Consequently, the selection of the appropriate model becomes paramount in ensuring the effectiveness of the recommendation system.

In the realm of cross-selling, the application of algorithms is instrumental in scrutinizing customer behavior and proposing complementary products [7]. The Apriori algorithm, a key player in association rule mining, adeptly identifies patterns in purchased items, enabling businesses to strategically promote related products and enhance the likelihood of cross-selling success [18]. Collaborative filtering, encompassing user–item interaction and item–item collaborative filtering, leverages collective preferences to provide personalized recommendations aligned with individual customer tastes, enriching the cross-selling potential [19].

Up-selling strategies, designed to promote higher-value products, rely on sophisticated algorithms like Decision Trees. These trees analyze customer behavior, pinpointing factors leading to premium purchases [20]. Concurrently, ranking models such as RankNet or LambdaMART contribute by predicting and ranking products based on their revenue potential, ensuring personalized recommendations that resonate with customer desires [21]. The synergy between Decision Trees and ranking models equips businesses to optimize up-selling initiatives and maximize revenue opportunities [20].

To craft effective bundle recommendations, businesses employ algorithms that comprehend inherent product relationships. Market Basket Analysis and clustering algorithms prove indispensable, revealing patterns in customer transactions and organizing products into cohesive clusters [18]. These insights empower businesses to recommend entire clusters as bundles, enhancing the overall shopping experience and boosting sales through well-matched product combinations.

Frequently-bought-together recommendations capitalize on advanced algorithms, with collaborative filtering playing a pivotal role. By analyzing user preferences and behaviors or drawing connections between similar items, this technique suggests products commonly purchased together [19]. Association rule mining complements this approach, identifying relationships within transaction data and offering insights into tandem purchases [22]. The amalgamation of collaborative filtering and association rule mining enhances recommendations, providing customers with suggestions aligned seamlessly with their preferences and purchase history.

In addition, personalized recommendations rely on a diverse set of algorithms, prominently featuring collaborative filtering to recommend products based on similar user profiles (user profiling) [23]. Content-based filtering suggests items based on the features of previously engaged-with products [24]. Matrix Factorization techniques, including SVD and ALS, contribute by breaking down user–item interaction matrices [25]. Deep Learning Models, such as RNNs and NCF, offer complex and accurate personalized recommendations by learning intricate patterns in user behavior data [26]. The fusion of these algorithms ensures a nuanced and highly personalized shopping experience, reflecting individual preferences and enhancing user satisfaction.

Crucially, many modern recommender systems adopt hybrid approaches, combining various techniques to harness their strengths for more accurate and diverse recommendations [27,28]. The choice of algorithms hinges on the dataset, product characteristics, and the specific goals of the e-commerce platform.

2.3. Natural Language Processing and Advanced Language Models

Natural language processing (NLP) stands as a pivotal force in the realm of artificial intelligence, empowering machines not only to understand but also to interpret and manipulate human language effectively [29]. It serves as the crucial link between human communication and machine comprehension, covering a diverse array of tasks, including speech recognition, text classification, sentiment analysis, machine translation, information extraction, recommender systems and question answering [30]. The techniques associated with NLP enable the extraction of meaningful insights from vast amounts of unstructured textual data, facilitating efficient information retrieval, analysis, and decision-making processes [31]. NLP finds diverse applications in the e-commerce domain, including the development of chatbots, virtual assistants, language translation services, content summarization tools, and sentiment analysis [32]. Through its capability to harness the power of language, NLP contributes to the creation of intelligent systems that understand and communicate with humans in a natural and intuitive manner.

2.4. The Evolution of Generative Pre-Trained Transformer (GPT) Models

Generative Pre-trained Transformer (GPT) models, built upon a foundational architecture, play a pivotal role in advancing NLP capabilities [33]. These models have significantly transformed the NLP landscape, demonstrating an exceptional ability to capture and comprehend intricate linguistic structures, context, and semantic nuances [34]. Through extensive pre-training on copious amounts of unlabeled textual data, GPT models gain a profound understanding of language patterns and relationships [30]. This pre-training equips GPT models to generate coherent and contextually relevant text, enhancing their overall effectiveness [29]. What distinguishes GPT models is their adaptability; they can be fine-tuned for specific NLP tasks, thereby enhancing their performance and applicability across various domains [35]. Excelling in diverse NLP applications, such as text generation, language translation, summarization, and content completion, GPT models elevate language comprehension, text generation quality, and the delivery of contextually relevant, personalized results to users [29]. The contributions of GPT models play a crucial role in expanding the horizons of NLP, fostering more intricate and efficient interactions between humans and intelligent systems [36].

3. Materials and Methods

The current research is centered on the implementation of cross-selling, up-selling, and personalized recommendation strategies for customers who have already added products to their basket. This encompasses the integration of recommender systems to enhance the overall shopping experience. The primary objective is to encourage customers to consider additional related products that complement those already in their cart. This strategic approach is designed to elevate customer satisfaction with the offered products and concurrently augment the overall order value and average order value (AOV), thereby fostering additional sales for the e-commerce platform. For the purpose of this research, three distinct unsupervised machine learning algorithms were chosen and trained on a shared dataset. Following their training, these algorithms were assigned the task of generating product recommendations for specific product titles and categories.

Each algorithm employed a unique approach, either clustering products or recommending items based on their attributes. The models aimed to provide recommendations or predictions for the most fitting product within the training set for each product in the testing set. The resultant recommendations for each product in the testing set, generated by the three trained models, were stored in a CSV file for subsequent analysis.

The selected algorithms were as follows:

- K-means clustering algorithm: It is a widely adopted unsupervised machine learning technique designed to organize data points into distinct groups or clusters, leveraging their shared characteristics [37]. The "K" in K-means signifies the algorithm's objective of identifying a specific number of clusters within the dataset [38]. Through iterative processes, the algorithm assigns data points to clusters and refines cluster centroids until a convergence criterion is satisfied.
- Content-based filtering (CBF): It stands as a distinctive recommendation system methodology, drawing its efficacy from the inherent attributes of items and user inclinations to curate personalized suggestions. Diverging from collaborative filtering, which relies on user–item interactions, CBF zeroes in on the intrinsic content characteristics of items and user profiles (user profiling) [19]. Items are delineated through features or metadata like keywords, genres, or pertinent descriptors. By scrutinizing users' historical preferences, the system adeptly suggests items aligning with their past choices or explicitly stated preferences [39].
- Hierarchical clustering: It is recognized as a robust technique in data analysis and clustering, orchestrating the arrangement of data into a hierarchical tree or dendrogram structure based on similarities among data points [40]. This method systematically builds clusters by iteratively merging or dividing existing clusters until a comprehensive hierarchy is established. The decision-making process in hierarchical clustering, deciding whether to group or separate data points, relies on a selected distance metric like the Euclidean distance or correlation [41]. The agglomerative approach starts with individual data points, progressively merging them into clusters, whereas the divisive approach begins with a single cluster and iteratively fragments it into smaller clusters [42]. This hierarchical representation provides a nuanced comprehension of relationships and structures within the dataset, unveiling insights into the diverse levels of similarity and dissimilarity between data points.

Post-prediction, the evaluation phase utilized the GPT-4 model, considering that the product suggested by the algorithms is related to the items already present in the customer's cart and could be recommended for combined purchase.

To fulfill the objectives of this paper, a specific methodology was adhered to, encompassing distinct steps for training, recommendation generation, recommendations' evaluation, and software development. The ensuing subsections outline these steps to present a comprehensive and cohesive strategy for the study.

3.1. Dataset Splitting and Preprocessing

For the research purposes, the Amazon UK Products dataset was utilized [43]. This specific dataset is available on the Kaggle platform and is distributed under the Open Data Commons Attribution License (ODC-By) v1.0. The ODC-By v1.0 license allows users to freely use, modify, and share datasets, provided proper attribution is given to the original data source. The dataset has a size of 137 MB and contains 2.2 million products. This dataset was chosen due to its origin from a reliable source, with each product in the collection offering comprehensive information such as a title, category, ASIN, price, reviews, stars, and more. In this study, although all columns were retained from the original dataset, only the title and categoryName columns were used for both training and recommendations.

In the initial phase, a random selection of 10,000 products was conducted using the train_test_split function from the sklearn library [44], chosen from the pool of 2.2 million products in the dataset. Subsequently, a new CSV dataset was generated to encompass this subset of products. Care was taken to ensure that for each entry in the new dataset, the fields of product title and category were populated. Subsequently, the dataset was split into training (70%), validation (15%), and test (15%) sets, and each set was saved into separate CSV files.

Both the 10,000-product dataset and the training, validation, and test sets are available in a GitHub repository [45].

3.2. Training K-Means Clustering, Content-Based Filtering (CBF), and Hierarchical Clustering Models for Product Recommendations

In this section, the methodology followed for both training and predictions/product recommendations is presented separately for each model. To expedite the results, the training and prediction phases were conducted on Google Colab using a A100 GPU [46]. The source code for each training and recommendation phase for the three algorithms is available in an ipynb file on GitHub [45].

3.2.1. K-Means Clustering Training

For the K-means clustering training phase, the process begins by importing the necessary libraries and mounting Google Drive to access and save files. Global variables for TF-IDF transformers are defined, and a function to transform input data using pre-trained TF-IDF transformers is implemented [47].

The K-means clustering model is trained using a combined dataset of training and validation data. The input features are extracted using TF-IDF for both the title and categoryName attributes. The number of clusters is set to 10, and the model is trained on the combined transformed data. The fine-tuned model, along with the TF-IDF transformers, is saved for future use.

The model's performance is evaluated on the validation set using the Silhouette Score, which measures how well defined the clusters in the data are. Recommendations are made based on the input data, with the CSV file being updated to include the kMeansRecommendation column.

To make recommendations, the model is loaded, and the input data is transformed using TF-IDF transformers. The cluster is predicted for the input data, and products within the same cluster are identified. The cosine similarity between the input product and each product in the cluster is calculated, and recommendations are sorted based on these similarity scores. The top N recommendations are then returned.

Finally, a function iterates through all the products in the test set. It retrieves recommendations, selects the top N recommendations for each product, and stores the first one in the kMeansRecommendation column.

3.2.2. Content-Based Filtering (CBF) Training

In the outlined procedure for training and generating product recommendations using content-based filtering (CBF), the initial step involves mounting Google Drive to access and save files. The Python script utilizes the pandas library to handle data frames, the joblib library for file I/O operations, and scikit-learn for natural language processing tasks, employing the TfidfVectorizer to transform product titles and categories into feature vectors [48]. A separate function is created to facilitate the transformation of input data based on pre-trained TF-IDF transformers, and the main training function is then employed to train these transformers using the provided training data.

The core recommendation process is encapsulated in a function that calculates the cosine similarity between the input product and those in the training data, sorts them by similarity scores, and returns a list of recommended product titles. Finally, another function utilizes the trained transformers to update a CSV file with recommendations, incorporating the cbfRecommendation column for each product. This function iterates through the existing data, applies the recommendation algorithm, and updates the CSV file accordingly. The entire process is orchestrated to enhance product recommendations based on the content-based filtering approach.

3.2.3. Hierarchical Clustering Training

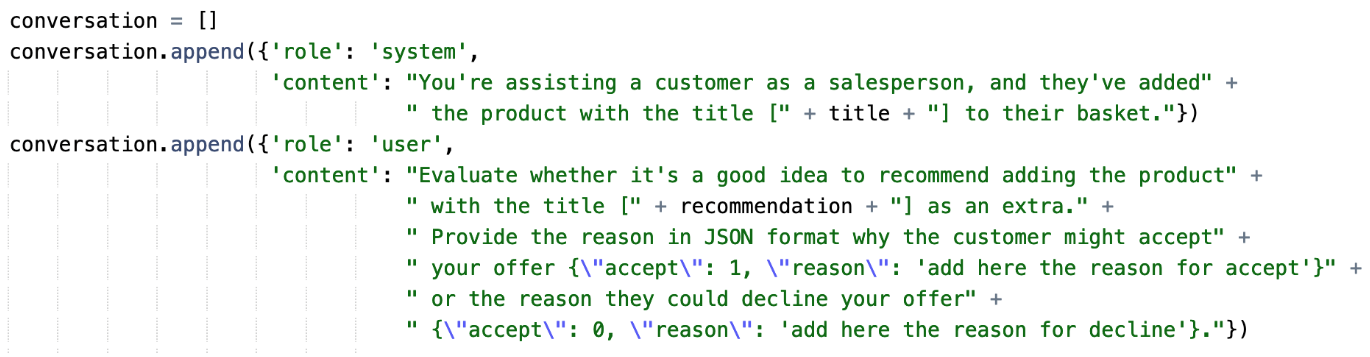
For hierarchical clustering, the initial steps involve importing the necessary libraries and mounting Google Drive to access and save files. Two global variables, title_tfidf and category_tfidf, are declared to store pre-trained TF-IDF transformers. The transform_input function is then defined to transform input data using these pre-trained transformers. The training process begins by combining the training and validation datasets and extracting features using TF-IDF for both the product title and category. The chosen number of clusters is set to 2, and an agglomerative clustering model is trained on the combined transformed data. The model, along with the TF-IDF transformers, is saved for future use. An evaluation is performed using the Silhouette Score.

To generate product recommendations, the recommend_product function is designed to predict the cluster for the input data, identify products in the same cluster, calculate cosine similarity scores, and sort products based on these scores. The top N recommended titles are then returned, and the first of them is selected. The get_recommendations function loads the pre-trained model, makes recommendations for a specific row in the dataset, and updates a CSV file with the recommendation information. The entire process is executed on a training and validation dataset, and the recommendations are stored in the hierarchicalRecommendation column in the test set.

3.3. Zero-Shot Evaluation Methodology Using GPT-4 Model for Assessing Product Recommendations

The primary objective of our research was to discover an innovative approach for evaluating the efficiency of the specific algorithms for product recommendation beyond the conventional methods presented in earlier studies. Many research endeavors comparing models in similar tasks often employ the similarity score and other evaluation metrics, which utilize the SequenceMatcher from the difflib module to calculate the similarity ratio between the product title and the recommended product title [49,50]. In this study, we propose the use of the GPT-4 model for evaluating the effectiveness of models for product recommendations. As mentioned in Section 2.3, the newly introduced GPT-4 model by OpenAI is a large language model that exhibits numerous applications in various domains and tasks.

At this stage of the research, we possessed a comprehensive CSV file containing product recommendations for each item in the test set. At this juncture, we constructed a prompt designed to prompt the GPT model to evaluate the recommendations made by the three models for each product in the test set. For every product and each recommendation, the GPT model was tasked with assessing the recommendation with a binary rating: 1 indicating a high likelihood of customer purchase, and 0 signifying a low probability of purchase. In addition to the binary evaluation, to gain further insights from the GPT model's assessment, the model was prompted to provide a textual justification for its decision. Both the evaluation scores and the accompanying textual justifications were stored in the CSV file. After multiple attempts, the selected prompt capable of simultaneously performing both tasks is presented in Figure 1.



```
conversation = []
conversation.append({'role': 'system',
                     'content': "You're assisting a customer as a salesperson, and they've added" +
                                " the product with the title [" + title + "] to their basket."})
conversation.append({'role': 'user',
                     'content': "Evaluate whether it's a good idea to recommend adding the product" +
                                " with the title [" + recommendation + "] as an extra." +
                                " Provide the reason in JSON format why the customer might accept" +
                                " your offer {\"accept\": 1, \"reason\": 'add here the reason for accept'}" +
                                " or the reason they could decline your offer" +
                                " {\"accept\": 0, \"reason\": 'add here the reason for decline'}."})
```

Figure 1. GPT prompt for evaluating product recommendations.

The dialogue in Figure 1 is structured as a series of dictionaries, with each dictionary corresponding to a turn in the conversation. The initial dictionary contains the system's prompt, functioning as a salesperson proposing a product recommendation aligned with an item already present in the customer's cart. The second dictionary contains the user's prompt, in which they ask the system to assess the product suggested by the unsupervised algorithms.

For the execution of the GPT-4 model, the official OpenAI API was utilized [51]. To achieve the desired JSON format for the output from the API, the prompt was carefully designed to provide clear instructions to the GPT model regarding the expected structure of the JSON format. In the majority of cases, the results were in the correct structure. However, for instances where the GPT model returned additional text, an additional function was implemented to locate the JSON within the text, convert it into a suitable dictionary, and appropriately process and store it in the CSV file.

3.4. Development of Flask-Based API Software for Integrated Training and Deployment of Product Recommendation Models

Our research aims to empower e-commerce owners by facilitating the integration of advanced machine learning algorithms into their recommendation systems, even without prior experience in machine learning or artificial intelligence. While there are existing automated tools or extensions for e-commerce platforms that offer recommendation system capabilities, some may rely on simplistic machine learning algorithms that lack customization for specific product sets. Moreover, these algorithms may not be easily trainable on low-budget web servers.

In contrast, our proposed solution, flask-based software, enables any webmaster with basic programming knowledge to deploy it on a web server (such as PythonAnywhere hosting). By leveraging the API features, the webmaster can seamlessly connect their e-commerce platform with our recommendation software. The distinct advantage lies in the ability to train the algorithm on the specific products of the e-commerce site, ensuring tailored product recommendations.

During the development phase, we consolidated functions previously used on Google Colab, organized them into classes, and encapsulated them within a flask application. Once set up on a web server, the webmaster can initiate the process by providing the e-commerce products' URL in a CSV format through the API. The software downloads the CSV file, conducts training and GPT evaluation, and identifies the algorithm that delivers the most effective product recommendations. Subsequently, the webmaster can utilize the same API to make calls, supplying a product title, and the software will predict the best-matching products likely to be purchased by the customer.

To further explain the API's deployment and interaction with the user, a flow chart was created using Lucidchart software [Figure 2]. This block diagram describes both the training phase and product recommendation phase.

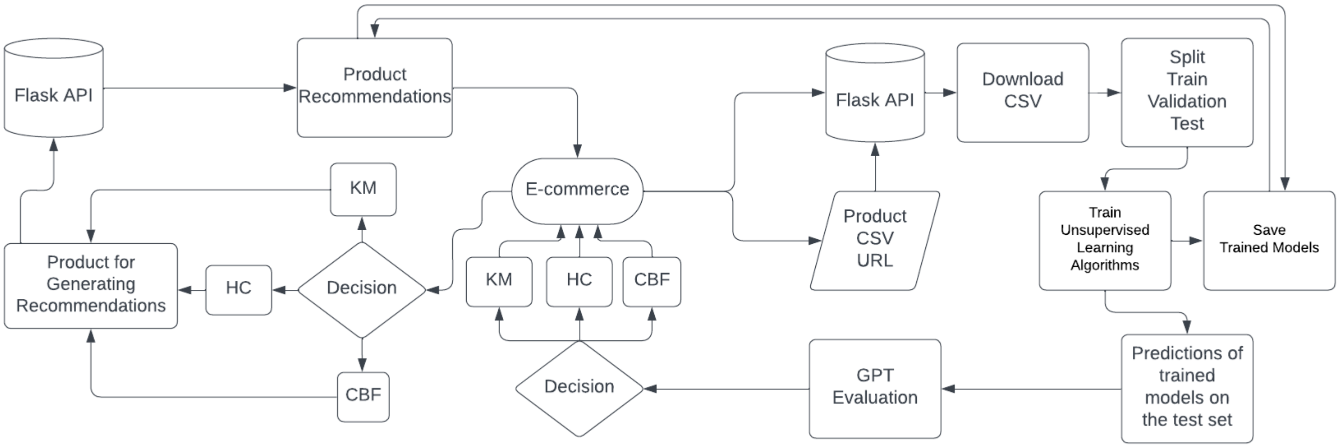

Flask API
Product Recommendations
Flask API
Download CSV
Split Train Validation Test
KM
E-commerce
Product CSV URL
Train Unsupervised Learning Algorithms
Save Trained Models
Product for Generating Recommendations
HC
Decision
KM
HC
CBF
CBF
Decision
GPT Evaluation
Predictions of trained models on the test set

Figure 2. Training and product recommendations using the flask API.

Our software can be further enhanced to provide additional functionalities for the webmaster. Moreover, expanding the training of algorithms for the recommendation task offers the webmaster a broader range of options, whether for prediction speed or improved recommendation outcomes. The source code of the software is available on GitHub in an open-source format under the MIT license [45].

3.5. Critical Factors Shaping Software Effectiveness in Collaborative and Real-Time Environments

In the realm of collaborative projects or real-time scenarios requiring human interaction, the effectiveness of the software discussed in this article hinges on several key factors:

- Seamless integration: The flask-based API offers an intuitive interface for integration, streamlining the process for developers and e-commerce proprietors to seamlessly incorporate and deploy the model in real-time applications.
- Automated training and evaluation: The software automates the training and evaluation processes, sparing users the intricacies of managing machine learning algorithms, model validation, and assessment.
- Enhanced natural language understanding with GPT-4: Leveraging GPT-4 for model evaluation harnesses its advanced natural language understanding capabilities, fostering effective human interaction. GPT-4's adeptness at refining the semantic context of product features has the potential to elevate the precision and relevance of product recommendations.
- Scalability and efficiency: The software stands out as an efficient and scalable solution, tailored to handle varying levels of structured data specific to e-commerce needs. This adaptability is crucial for collaborative projects and real-time scenarios.
- User-centric experience: In real-time scenarios, success often hinges on the user experience. The model generates personalized and effective recommendations, enhancing the likelihood of positive user reception.

While the proposed software model exhibits promising features for collaborative projects and real-time scenarios, particularly in the context of e-commerce, practical implementation and user feedback are indispensable for a comprehensive assessment of its performance in such environments.

4. Research Results

In the preceding sections, we introduced three distinct unsupervised learning algorithms tailored to product recommendations, with a specific focus on in-cart product recommendations. Through training on both a given training set and a validation set, these algorithms were meticulously trained to provide optimal predictions and recommendations for the test set.

Building upon this, we propose an innovative approach in which the GPT-4 model takes on the role of evaluating the effectiveness of the recommendations generated by the trained models. This evaluation involves assigning a binary rating along with a contextual justification for the given rating. The subsequent section unveils the outcomes of this evaluation, addressing the research queries outlined in the Section 1.

4.1. Comparing Unsupervised Learning Algorithms Trained for Product Recommendation in E-Commerce

- Research Question 1: Which unsupervised recommender system algorithm demonstrates superior efficacy in product recommendation tasks?
- Research Statement 1: The content-based filtering (CBF) and k-means clustering-trained models demonstrate higher accuracy in in-cart product recommendation tasks.

In accordance with Section 3.2, three unsupervised models underwent training on a shared dataset consisting of 7000 products, validated with a set of 1500 items. Following the training phase, the models were tasked with generating product recommendations from the training set for each of the 1500 products in the test set. The resulting predictions were recorded in a CSV file, and the evaluation process was handed over to the GPT-4 model. The GPT-4 model evaluated the recommendations, assigning a binary rating on a scale from 0 to 1. A rating of 1 indicated a high likelihood of customer purchase based on the recommendation, while a rating of 0 signified a low probability of purchase. It is crucial to note that the focus of this evaluation was on in-cart product recommendations, where the customer already has at least one product in their basket, and we propose supplementary products likely to be added to their cart. The outcomes of the GPT-4 model evaluations are depicted in Figure 3 and Table 1.

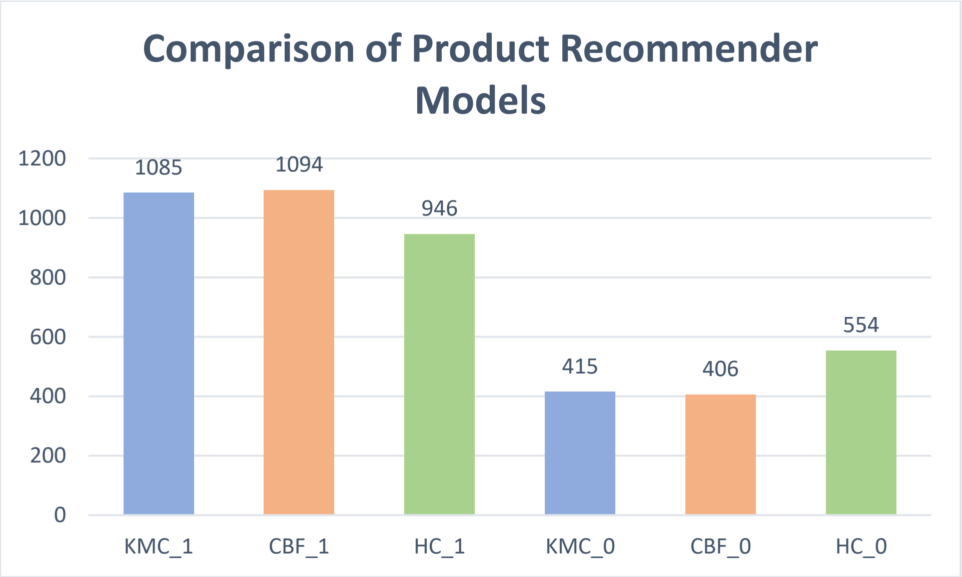


Figure 3. GPT-4 evaluations of unsupervised learning-trained models.

Table 1. Efficiency of unsupervised learning-trained models in recommender systems.

	0	1	Total
KMC Evaluation	415	1085	1500
CBF Evaluation	406	1094	1500
HC Evaluation	554	946	1500

Figure 3 reveals that the content-based filtering (CBF)-trained model outperforms the others, achieving a commendable 72.93% accuracy, equivalent to 1094 correct recommendations out of the 1500 products assessed by the GPT model. Following closely, the k-means clustering-trained model achieved 72.33% accuracy, providing 1085 correct recommendations for the products in the test set. In contrast, the hierarchical clustering-trained model exhibited a lower success rate, accurately recommending 946 out of the 1500 test set products, corresponding to 63.07% accuracy.

The GPT model evaluation underscores the superior performance of the content-based filtering (CBF)-trained model, likely attributed to its more sophisticated nature. The k-means clustering-trained model closely follows, with a negligible distance from CBF, indicating high accuracy in its recommendations.

It is noteworthy to emphasize that achieving an accuracy of 70% is considered significant in recommendation tasks.

4.2. Evaluation of Models Using the GPT-4 Model

- Research Question 2: Is the GPT-4 model capable of evaluating the predictions of traditional unsupervised models?
- Research Statement 2: The GPT-4 model exhibits significant potential for evaluating predictions from unsupervised models.

Upon delving into our research findings, it is imperative to highlight the effectiveness of the GPT-4 model as an evaluation tool. This model not only assesses trained models in a binary format but also provides insightful justifications for its ratings. Table 2 presents some of the ratings assigned by the GPT-4 model along with their corresponding justifications.

Table 2. In-cart product recommendations and GPT model evaluation.

In-Cart Product	Recommendation Product	Model	Binary Rating	Justification
Holibanna Human Skull Teaching Model Lifesize Human skeleton Anatomical Head Bone Model Halloween Decoration	267Pcs Molecular Model Organic Inorganic Structure Kit, Atoms Bonds Orbitals Chemistry Structure Kit Atom Link Model, Educational Molecule Set for Teachers Students Student Chemistry Learning	HC	0	The customer is purchasing a human skull teaching model for Halloween decoration, which suggests they are interested in anatomical models rather than chemistry models. Recommending a molecular model kit might not align with their current purchase intention.
Hellyhansen Women’s Crew Sweatshirt Women’s Sweatshirt	VOIO Men’s Hoodies, Warm Sweatshirt Crew Neck Print Sweatshirt Christmas Tree Plaid Print Sweatshirt 3D Pullover Suitable Autumn Winter Sweatshirt	CBF	0	The recommended product is for men, while the customer has added a women’s sweatshirt to their basket. It’s unlikely that the customer would be interested in purchasing a product that is not suitable for their gender.
2 Pack Silicone Foldable Cups BPA-Free Collapsible Travel Mugs 270ML Portable Reusable Cup Set with Plastic Sealing Lid for Outdoor Picnic Camping Hiking Traveling (Grey & Blue)	Light my Fire Camping Mess Kit—Picnic Set 6 Pcs—Lunchkit—Camp Mess Kit—Reusable Camping Plates and Bowls Set—BPA Free—Microwave & Dishwasher—Camping Picnic Set	KMC	1	The customer might accept adding the Light my Fire Camping Mess Kit because it complements the Silicone Foldable Cups. The Mess Kit includes reusable camping plates and bowls, which would be useful for outdoor picnics, camping, and hiking. Additionally, the fact that it is microwave and dishwasher safe makes it convenient for the customer to use and clean. Overall, the Camping Mess Kit enhances the customer’s outdoor dining experience and provides them with a complete set of cookware.

KMC: K-means clustering; HC: hierarchical clustering; CBF: content-based filtering.

Upon examining Table 2, it becomes evident that the GPT-4 model possesses significant potential as an evaluation tool. Its unique capabilities extend beyond accurately assessing predictions from other models; as an NLP model, it can also articulate justifications, effectively putting itself in the customer's shoes. By adopting our novel approach, data analysts can enhance their unsupervised recommender system algorithms, enabling more precise and insightful predictions.

4.3. Comparison of LLM Evaluation and Human Evaluation

- Research Question 3: Can LLMs replace human evaluations?
- Research Statement 3: LLMs present a more cost-effective and time-efficient alternative to human evaluations, but they cannot surpass human prowess.

This research inquiry leans more towards theory than experimentation. Natural language processing models, including large language models (LLMs), exhibit the capability to comprehend the meaning behind human-provided text through context, and they can make evaluations akin to humans. However, human evaluations stem not only from context but also from life experiences, a dimension challenging for LLMs to replicate. Therefore, while LLMs can be utilized for evaluations in product recommendation tasks, they do not surpass humans in evaluative prowess.

Human evaluations, though effective, come with a high cost, and it is presumed that the human, often the webmaster in this context, possesses unlimited time for the task. In contrast, LLM evaluations can offer a more cost-effective and time-efficient alternative. In conclusion, LLMs may not match the success rate of human evaluations, but they can execute similar tasks swiftly, saving both time and money for the webmaster.

4.4. Natural Language Model Evaluations on Unsupervised Learning Algorithms

- Research Question 4: Do NLPs have the capability to evaluate recommendations from unsupervised learning algorithms?
- Research Statement 4: NLPs, specifically LLMs like the GPT-4 model, have proven through our study to be effective tools for evaluating product recommendations generated by unsupervised learning models.

Based on our research findings, it is evident that NLP models, particularly the GPT-4 model, demonstrate remarkable capabilities in performing diverse tasks, including the evaluation of product recommendations generated by unsupervised models. These evaluations exhibit high accuracy in both binary ratings and justifications.

A key question that arose during our research is why we persist in using unsupervised recommender system algorithms like clustering and CBF for product recommendation tasks instead of leveraging an LLM for potentially more efficient results. While an LLM might indeed excel at a comparable task, it is crucial to consider that incorporating an LLM introduces costs, whether in terms of hardware or API expenses. In contrast, evaluating the unsupervised recommender system algorithms presented in this study incurs minimal hardware costs for product recommendations.

Furthermore, our study proposes utilizing the evaluation capabilities of GPT-4 not only for providing ratings but also to discern which product recommendation algorithm is more efficient. Data scientists can leverage this by multi-training and evaluating these models, adjusting parameters, or using hyperparameter tuning to enhance their precision, potentially exceeding the 70% accuracy reported in this study.

4.5. The Contribution of the Software to E-Commerce

- Research Question 5: Can the proposed flask-based API contribute to the accessibility and usability of sophisticated recommendation systems for e-commerce owners?
- Research Statement 5: The software proposed in this study has the potential to enhance both the product recommendation capabilities of e-commerce and the resulting benefits from additional sales, thereby providing high accessibility for e-commerce owners aiming to improve their recommendation system and, consequently, their e-commerce sales.

The resolution to our final research question lies in the hands of e-commerce owners who opt to employ our software to elevate their e-commerce platforms' product recommendation capabilities. The effectiveness of our software can be gauged by the additional sales potential it unlocks for e-commerce businesses. This sophisticated software not only saves time and is cost-efficient but also proves highly effective, combining the insights of an LLM with the sophisticated nature of unsupervised recommender system algorithms. Undoubtedly, our proposal stands to benefit e-commerce owners, whether by saving time on manually adding product recommendations or by enhancing the effectiveness of in-cart product recommendations.

5. Recommender System Evaluation and Discussion

In Section 4, the research outcomes are presented, addressing the research questions and extracting valuable research statements. This section is dedicated to assessing the recommender systems developed in this research through a traditional approach. The outcomes of this evaluation will be juxtaposed with those of the GPT-powered evaluation tool proposed in this article, shedding light on how our assessment tool attains its binary ratings and how closely its evaluations align with the traditional approach. Following this, a discussion ensues regarding the limitations and inherent risks associated with our proposed tool.

5.1. Assessing Recommender Systems' Effectiveness

Assessing the effectiveness of a recommender system involves gauging how well the recommended products align with user preferences or actual behavior. Various evaluation metrics, such as user surveys, accuracy metrics, user engagement metrics, precision and recall, A/B testing, and more, are commonly used. However, most of these methods require human evaluation. For instance, user surveys necessitate asking users about their satisfaction with recommended products, accuracy metrics require user ratings for recommended products, user engagement metrics involve tracking click-through rates and time spent on recommended product pages, precision and recall require feedback in binary form, and A/B testing involves randomly selected human evaluators assessing product recommendations.

Our proposed methodology introduces an alternative using GPT-4 as an evaluation tool. GPT-4 acts as a virtual human evaluator, assigning binary ratings based on its assessment of whether a product recommendation from an unsupervised model is more likely (1) or not likely (0) to be purchased by the customer. While the ratings are binary, the precision and recall technique cannot be applied due to its requirement of two columns—one with true values and one with predicted values.

The most effective means of evaluating both the recommendations of the unsupervised models and the performance of the GPT-4-powered evaluation tool is undoubtedly through a human evaluation. However, with 1500 rows of test data and 7000 rows of training data, conducting a human evaluation becomes time-consuming. An alternative is to assess the precision without human interference by calculating the similarity score between the product title and the recommended product title. This score possesses the ability to evaluate recommendations produced by unsupervised trained models while concurrently assessing the effectiveness of evaluations conducted by the GPT-4 model. Moreover, the similarity score can be considered a practical alternative to the GPT-4-powered evaluation tool suggested in this research.

For our study, we employed the cosine similarity measure to evaluate the effectiveness of our recommender systems. This methodology uses the TfidfVectorizer to convert the text into TF-IDF vectors and then calculates the cosine similarity between these vectors.

The similarity scores were calculated between the title and the recommendations made by the unsupervised trained models, and the results were stored in different columns for each of our recommendation systems (K-means clustering, hierarchical clustering, and content-based filtering). Descriptive statistics for each new column have been computed and are showcased in Table 3.

Table 3. Descriptive statistics for the similarity scores for each recommendation system.

Metrics	KMC	HC	CBF
count	1500	1500	1500
mean	0.315991	0.228553	0.304138
std	0.203004	0.163804	0.200361
min	0.000000	0.000000	0.000000
25%	0.151775	0.103950	0.145600
50%	0.276650	0.190900	0.259550
75%	0.440075	0.310575	0.426925
max	1.000000	0.958000	1.000000

KMC: K-means clustering; HC: hierarchical clustering; CBF: content-based filtering.

Interpreting the findings from the descriptive statistics presented in Table 3, all three recommendation systems consistently display a count of 1500, indicating that similarity scores were calculated for the entire dataset. The mean similarity scores for the K-means clustering (KMC), content-based filtering (CBF), and hierarchical clustering (HC) recommendation systems are 0.316, 0.304, and 0.229, respectively. Higher mean values suggest that, on average, the recommendations moderately resemble the original titles.

Examining standard deviations, KMC and CBF exhibit similar values (around 0.203 and 0.200, respectively), implying moderate variability in scores, while HC shows a slightly lower standard deviation (around 0.164), suggesting relatively less variability. The minimum similarity scores are 0 for all the systems, indicating instances with no similarity between the titles and the recommendations, while maximum scores of 1 suggest perfect alignment in certain cases.

Based on the descriptive analysis, it is evident that KMC and CBF perform similarly and significantly outperform HC. As per Table 1 and Figure 3, our GPT-powered evaluation tool indicates that CBF and KMC are closely aligned in terms of precision, with approximately 72.93% for CBF and 72.33% for KMC. In contrast, the HC recommendation system exhibits a lower precision of about 63.07%. These results suggest that our proposed methodology closely aligns with traditional approaches, such as the use of similarity scores.

In our evaluation methodology, CBF is slightly more precise in its recommendations than KMC by about 0.6%. Using similarity scores, KMC appears slightly more precise than CBF by 1.2%. Both evaluation tools underscore that the HC recommendation system is significantly less precise, with CBF and KMC consistently recommending the best products, which are more likely to be purchased by the customer.

As mentioned earlier, the evaluation methodology proposed in this article closely aligns with traditional approaches like the use of similarity scores to determine which model has the greatest potential for generating product recommendations that are likely to be purchased by a customer.

Upon reviewing Table 3, it becomes evident that the similarity scores appear to be quite low. It is important to note that a score of 1 indicates perfect alignment, while 0 signifies no alignment. In contrast, our proposed methodology demonstrates success percentages ranging from 63.07% to 72.93%. This observation highlights significant disparities in how the evaluation tools assess recommendations.

Unlike similarity scores, which attempt to measure the likeness between a given product and its recommended counterpart in vector terms, our GPT-4 evaluation tool takes a different approach. Instead of evaluating textual similarity, it comprehends the meaning behind each product title and recommendation within the context, leading to a more thoughtful evaluation.

From a customer perspective, a recommendation that is perfectly aligned with the product already in their cart may not be the optimal suggestion. Rather, understanding the nuanced meaning of the product title and proposing a complementary product, even if not perfectly aligned, proves to be a superior recommendation. This approach holds greater potential for enticing customers to make a purchase by offering supplementary products that enhance their overall shopping experience.

The primary objective of this article is to introduce an alternative evaluation tool for recommender systems, highlighting the effectiveness of large language models (LLMs), particularly the GPT-4 model. Undoubtedly, to comprehensively assess whether our proposed methodology surpasses traditional approaches, human evaluation is deemed necessary. It is crucial to note, however, that the understanding capabilities of LLMs hold potential across various domains, making them a versatile evaluation tool applicable in diverse contexts and industries.

5.2. Threats to Validity

The envisioned incorporation of GPT-4 as an evaluation tool for assessing unsupervised learning algorithms in recommender systems holds promise, but its implementation comes with inherent limitations and risks.

- Data bias and generalization issues: The efficacy of the model hinges on the quality and diversity of the training data. Biased or insufficiently varied data provided by e-commerce owners may result in skewed recommendations that lack generalizability to a broader audience.
- Overfitting: The integration of multiple models, particularly with a substantial language model like GPT-4, raises the risk of overfitting. This occurs when the model excels on training data but struggles to generalize to new, unseen data.
- Computational resources: The use of the GPT-4 model in the proposed recommender system could incur significant API costs for both fine-tuning and predictions, depending on the dataset size. This presents challenges for smaller e-commerce platforms with limited budgets.
- Interpretability: Complex integrated models may lack interpretability, making it difficult to comprehend how the model generates specific recommendations. This lack of transparency can be a concern for businesses that prioritize understanding the decision-making process.
- Privacy concerns: The advanced natural language understanding capabilities bring forth privacy concerns, especially if sensitive or personal information is inadvertently captured and utilized in the recommendation process. The software developed relies on trust in the data provided by e-commerce owners through the API without evaluating potential sensitivity.
- Maintenance and updates: Keeping the recommendation system up-to-date with the latest data and ensuring compatibility with evolving e-commerce platforms pose a challenge. Regular maintenance and updates are imperative, considering the possibility of GPT-4 being deprecated in the future or changes in API functions and Python libraries.
- User engagement and satisfaction: While the experimental results may suggest the effectiveness of the proposed methodology, it is essential to acknowledge that the GPT-powered evaluation tool comprehends the meaning behind each word, diverging from reliance on simple word similarity for its evaluation outcomes. Consequently, it becomes imperative to factor in user satisfaction and engagement. Users may not consistently prioritize the utmost accuracy in recommendations, underscoring the importance of striking a balance between precision and accommodating user preferences.

In conclusion, while the proposed integration of GPT-4 as an evaluation tool signifies a groundbreaking approach, addressing and mitigating these potential limitations and risks is vital for its successful real-world implementation.

6. Conclusions

In summary, this paper presents an innovative methodology that integrates unsupervised learning models, such as K-means clustering, content-based filtering (CBF), and hierarchical clustering, with the cutting-edge GPT-4 language model (LLM) to elevate e-commerce product recommendation systems. The training of these models, utilizing both training and validation sets, involves generating recommendations for a designated test set. To assess the models' recommendations, the GPT model assigns ratings ranging from 0 to 1, where 1 signifies a high likelihood of customer purchase based on the recommendation, and 0 indicates a low probability of purchase. Notably, the GPT model also provides justifications for its ratings within the same prompt.

The study's outcomes reveal compelling results, with the CBF-trained model achieving an accuracy of 72.93%, the K-means clustering-trained model achieving 72.33%, and the hierarchical clustering-trained model achieving 63.07% in the product recommendation evaluations conducted by the GPT model. Additionally, it is observed that LLMs, particularly the GPT-4 model, can effectively evaluate recommender system algorithms' predictions. Consequently, the central proposition of this study advocates for the utilization of LLMs as an evaluation tool, especially in the context of product recommendation tasks. While acknowledging that GPT models cannot replace human intelligence in evaluation tasks, their demonstrated effectiveness underscores their valuable contribution, particularly in the e-commerce domain.

To translate this knowledge into practical application, we developed user-friendly flask-based software with an easily installable API. This tool is designed to support e-commerce owners in enhancing their product recommendation systems, consolidating the synergy of unsupervised recommender system algorithms and the advanced language understanding offered by GPT-4 LLM, thereby providing an actionable solution for businesses seeking to optimize their customer engagement strategies.

Author Contributions: Conceptualization, K.I.R. and N.D.T.; methodology, K.I.R., N.D.T. and D.K.N.; software, K.I.R.; validation, K.I.R. and N.D.T.; formal analysis, K.I.R., N.D.T. and D.K.N.; investigation, K.I.R., N.D.T. and D.K.N.; resources, K.I.R.; data curation, K.I.R.; writing—original draft preparation, K.I.R. and N.D.T.; writing—review and editing, K.I.R., N.D.T. and D.K.N.; visualization, K.I.R.; supervision, D.K.N. and N.D.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data supporting the reported results can be found at ref. [45].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Agarwal, A.; Chakraborty, M.; Chowdary, C.R. Does Order Matter? Effect of Order in Group Recommendation. *Expert Syst. Appl.* **2017**, *82*, 115–127. [CrossRef]
2. Kalyan, K.S. A Survey of GPT-3 Family Large Language Models Including ChatGPT and GPT-4. *Nat. Lang. Process. J.* **2023**, *6*, 100048. [CrossRef]
3. Products Suggestions for WooCommerce—WordPress Plugin | WordPress.Org. Available online: https://wordpress.org/plugins/cart-products-suggestions-for-woocommerce/ (accessed on 18 February 2024).
4. Lai, C.H.; Tseng, K.C. Applying Deep Learning Models to Analyze Users' Aspects, Sentiment, and Semantic Features for Product Recommendation. *Appl. Sci.* **2022**, *12*, 2118. [CrossRef]
5. Li, B.; Xu, H.; Zhao, Q.; Su, P.; Chabbi, M.; Jiao, S.; Liu, X. OJXPERF: Featherlight Object Replica Detection for Java Programs. In Proceedings of the 44th International Conference on Software Engineering 2022, Pittsburgh, PA, USA, 21–29 May 2022; pp. 1558–1570. [CrossRef]
6. Li, B.; Zhao, Q.; Jiao, S.; Liu, X. DroidPerf: Profiling Memory Objects on Android Devices. In Proceedings of the Annual International Conference on Mobile Computing and Networking, Madrid, Spain, 2–6 October 2023; pp. 75–89. [CrossRef]
7. Hell, F.; Taha, Y.; Hinz, G.; Heibei, S.; Müller, H.; Knoll, A. Graph Convolutional Neural Network for a Pharmacy Cross-Selling Recommender System. *Information* **2020**, *11*, 525. [CrossRef]
8. Ghoshal, A.; Mookerjee, V.S.; Sarkar, S. Recommendations and Cross-Selling: Pricing Strategies When Personalizing Firms Cross-Sell. *J. Manag. Inf. Syst.* **2021**, *38*, 430–456. [CrossRef]

9. Vatavwala, S.; Kumar, B.; Sharma, A. Enhancing Upselling and Cross-Selling in Business-to-Business Markets: The Critical Need to Integrate Customer Service and Sales Functions. In *Customer Centric Support Services in the Digital Age*; Palgrave Macmillan: Cham, Switzerland, 2024; pp. 199–216. [CrossRef]
10. Lesage, L.; Deaconu, M.; Lejay, A.; Meira, J.A.; Nichil, G.; State, R. A Recommendation System for Car Insurance. *Eur. Actuar. J.* **2020**, *10*, 377–398. [CrossRef]
11. Park, C.H.; Yoon, T.J. The Dark Side of Up-Selling Promotions: Evidence from an Analysis of Cross-Brand Purchase Behavior. *J. Retail.* **2022**, *98*, 647–666. [CrossRef]
12. Zhu, T.; Harrington, P.; Li, J.; Tang, L. Bundle Recommendation in ECommerce. In Proceedings of the SIGIR 2014—37th International ACM SIGIR Conference on Research and Development in Information Retrieval, Gold Coast, Australia, 6–11 July 2014; pp. 657–666. [CrossRef]
13. Zhou, H.; Xiong, F.; Chen, H.A.; Zhou, H.; Xiong, F.; Chen, H. A Comprehensive Survey of Recommender Systems Based on Deep Learning. *Appl. Sci.* **2023**, *13*, 11378. [CrossRef]
14. Alcaraz-Herrera, H.; Cartlidge, J.; Toumpakari, Z.; Western, M.; Palomares, I. EvoRecSys: Evolutionary Framework for Health and Well-Being Recommender Systems. *User Model. User-Adapt. Interact.* **2022**, *32*, 883–921. [CrossRef]
15. Chen, J.; Chen, W.; Huang, J.; Fang, J.; Li, Z.; Liu, A.; Zhao, L. Co-Purchaser Recommendation for Online Group Buying. *Data Sci. Eng.* **2020**, *5*, 280–292. [CrossRef]
16. Stöckli, D.R.; Khobzi, H. Recommendation Systems and Convergence of Online Reviews: The Type of Product Network Matters! *Decis. Support Syst.* **2021**, *142*, 113475. [CrossRef]
17. Wijaya, I.W.R. Mudjahidin Development of Conceptual Model to Increase Customer Interest Using Recommendation System in E-Commerce. *Procedia Comput. Sci.* **2022**, *197*, 727–733. [CrossRef]
18. Aldino, A.A.; Pratiwi, E.D.; Setiawansyah; Sintaro, S.; Putra, A.D. Comparison of Market Basket Analysis to Determine Consumer Purchasing Patterns Using Fp-Growth and Apriori Algorithm. In Proceedings of the 2021 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), Banyuwangi, Indonesia, 27–28 October 2021; pp. 29–34. [CrossRef]
19. Tewari, A.S. Generating Items Recommendations by Fusing Content and User-Item Based Collaborative Filtering. *Procedia Comput. Sci.* **2020**, *167*, 1934–1940. [CrossRef]
20. Rahmatillah, I.; Astuty, E.; Sudirman, I.D. An Improved Decision Tree Model for Forecasting Consumer Decision in a Medium Groceries Store. In Proceedings of the 2023 IEEE 17th International Conference on Industrial and Information Systems, ICIIS, Peradeniya, Sri Lanka, 25–26 August 2023; pp. 245–250. [CrossRef]
21. Jbene, M.; Tigani, S.; Rachid, S.; Chehri, A. Deep Neural Network and Boosting Based Hybrid Quality Ranking for E-Commerce Product Search. *Big Data Cogn. Comput.* **2021**, *5*, 35. [CrossRef]
22. Telikani, A.; Gandomi, A.H.; Shahbahrami, A. A Survey of Evolutionary Computation for Association Rule Mining. *Inf. Sci.* **2020**, *524*, 318–352. [CrossRef]
23. Wang, K.; Zhang, T.; Xue, T.; Lu, Y.; Na, S.G. E-Commerce Personalized Recommendation Analysis by Deeply-Learned Clustering. *J. Vis. Commun. Image Represent.* **2020**, *71*, 102735. [CrossRef]
24. Javed, U.; Javed, U.; Shaukat, K.; Hameed, I.A.; Iqbal, F.; Alam, T.M.; Luo, S. A Review of Content-Based and Context-Based Recommendation Systems. *Int. J. Emerg. Technol. Learn. iJET* **2021**, *16*, 274–306. [CrossRef]
25. Isinkaye, F.O. Matrix Factorization in Recommender Systems: Algorithms, Applications, and Peculiar Challenges. *IETE J. Res.* **2023**, *69*, 6087–6100. [CrossRef]
26. Wu, G.; Sanner, S.; Luo, K.; Soh, H. Deep Language-Based Critiquing for Recommender Systems. In Proceedings of the RecSys 2019—13th ACM Conference on Recommender Systems, Copenhagen, Denmark, 16–20 September 2019; pp. 137–145. [CrossRef]
27. Bhaskaran, S.; Marappan, R.; Santhi, B. Design and Analysis of a Cluster-Based Intelligent Hybrid Recommendation System for E-Learning Applications. *Mathematics* **2021**, *9*, 197. [CrossRef]
28. Ko, H.; Lee, S.; Park, Y.; Choi, A. A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. *Electronics* **2022**, *11*, 141. [CrossRef]
29. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. LLMs in E-Commerce: A Comparative Analysis of GPT and LLaMA Models in Product Review Evaluation. *Nat. Lang. Process. J.* **2024**, *6*, 100056. [CrossRef]
30. Rothman, D. *Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT*; Packt Publishing: Birmingham, UK, 2021.
31. Li, B.; Hou, Y.; Che, W. Data Augmentation Approaches in Natural Language Processing: A Survey. *AI Open* **2022**, *3*, 71–90. [CrossRef]
32. Fanni, S.C.; Febi, M.; Aghakhanyan, G.; Neri, E. Natural Language Processing. In *Introduction to Artificial Intelligence*; Springer: Cham, Switzerland, 2023; pp. 87–99. [CrossRef]
33. de Curtò, J.; de Zarzà, I.; Calafate, C.T. Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles. *Drones* **2023**, *7*, 114. [CrossRef]
34. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Unveiling Sustainability in Ecommerce: GPT-Powered Software for Identifying Sustainable Product Features. *Sustainability* **2023**, *15*, 12015. [CrossRef]
35. Liu, Z.; Yu, X.; Zhang, L.; Wu, Z.; Cao, C.; Dai, H.; Zhao, L.; Liu, W.; Shen, D.; Li, Q.; et al. DeID-GPT: Zero-Shot Medical Text De-Identification by GPT-4. *arXiv* **2023**, arXiv:2303.11032.

36. Zhang, M.; Li, J. A Commentary of GPT-3 in MIT Technology Review 2021. *Fundam. Res.* **2021**, *1*, 831–833. [CrossRef]
37. Bandyopadhyay, S.; Thakur, S.S.; Mandal, J.K. Product Recommendation for E-Commerce Business by Applying Principal Component Analysis (PCA) and K-Means Clustering: Benefit for the Society. *Innov. Syst. Softw. Eng.* **2021**, *17*, 45–52. [CrossRef]
38. Sinaga, K.P.; Yang, M.S. Unsupervised K-Means Clustering Algorithm. *IEEE Access* **2020**, *8*, 80716–80727. [CrossRef]
39. Nguyen, L.V.; Nguyen, T.H.; Jung, J.J. Content-Based Collaborative Filtering Using Word Embedding: A Case Study on Movie Recommendation. In Proceedings of the RACS'20: International Conference on Research in Adaptive and Convergent Systems, Gwangju, Republic of Korea, 13–16 October 2020; pp. 96–100. [CrossRef]
40. Sangaiah, A.K.; Javadpour, A.; Ja'fari, F.; Zhang, W.; Khaniabadi, S.M. Hierarchical Clustering Based on Dendrogram in Sustainable Transportation Systems. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 15724–15739. [CrossRef]
41. Lamb, D.S.; Downs, J.; Reader, S. Space-Time Hierarchical Clustering for Identifying Clusters in Spatiotemporal Point Data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 85. [CrossRef]
42. Li, T.; Rezaeipanah, A.; Tag El Din, E.S.M. An Ensemble Agglomerative Hierarchical Clustering Algorithm Based on Clusters Clustering Technique and the Novel Similarity Measurement. *J. King Saud Univ.—Comput. Inf. Sci.* **2022**, *34*, 3828–3842. [CrossRef]
43. Amazon UK Products Dataset 2023 (2.2M Products). Available online: https://www.kaggle.com/datasets/asaniczka/amazon-uk-products-dataset-2023 (accessed on 3 February 2024).
44. Sklearn.Model_Selection.Train_Test_Split—Scikit-Learn 1.4.0 Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed on 3 February 2024).
45. GitHub—Kroumeliotis/Product-Recommendations-Software-Unsupervised-Models-Evaluated-by-GPT-4-LLM: Precision-Driven Product Recommendations Software: Un-Leashing the Power of Unsupervised Models, Evaluated by GPT-4 LLM for Enhanced Recommender Systems. Available online: https://github.com/kroumeliotis/Product-Recommendations-Software-Unsupervised-Models-Evaluated-by-GPT-4-LLM (accessed on 3 February 2024).
46. Colab.Google. Available online: https://colab.google/ (accessed on 3 February 2024).
47. Walkowiak, T. Subject Classification of Texts in Polish—From TF-IDF to Transformers. In Proceedings of the Sixteenth International Conference on Dependability of Computer Systems DepCoS-RELCOMEX, Wrocław, Poland, 28 June–2 July 2021; pp. 457–465. [CrossRef]
48. Kumar, V.; Subba, B. A Tfidfvectorizer and SVM Based Sentiment Analysis Framework for Text Data Corpus. In Proceedings of the 26th National Conference on Communications (NCC), Kharagpur, India, 21–23 February 2020. [CrossRef]
49. Hong, Y.; Tantithamthavorn, C.; Thongtanunam, P.; Aleti, A. CommentFinder: A Simpler, Faster, More Accurate Code Review Comments Recommendation. In Proceedings of the ESEC/FSE 2022—30th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Singapore, 14–18 November 2022; pp. 507–519. [CrossRef]
50. Hasani, A.M.; Singh, S.; Zahergivar, A.; Ryan, B.; Nethala, D.; Bravomontenegro, G.; Mendhiratta, N.; Ball, M.; Farhadi, F.; Malayeri, A. Evaluating the Performance of Generative Pre-Trained Transformer-4 (GPT-4) in Standardizing Radiology Reports. *Eur. Radiol.* **2023**, 1–9. [CrossRef] [PubMed]
51. OpenAI API. Available online: https://openai.com/blog/openai-api (accessed on 14 November 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.