Data Imputation using Large Language Model to Accelerate Recommender System

Zhicheng Ding
zhicheng.ding@columbia.edu
Columbia University
New York, NY, USA

Jiahao Tian
jtian83@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Zhenkai Wang
kay.zhenkai.wang@utexas.edu
The University of Texas at Austin
Austin, Texas, USA

Jinman Zhao
jinman.zhao@mail.utoronto.ca
Univeristy of Toronto
Toronto, Ontario, Canada

Siyang Li
lisiyang98@hotmail.com
Pace University
New York, NY, USA

Abstract

The importance of recommender systems continues to grow as the volume of data generated increases. A robust recommender system can significantly enhance user experience and engagement. However, these systems often face challenges due to missing data, which can arise from various factors, including user privacy concerns and other reasons. In this paper, we propose a framework to address the challenge of sparse and missing data in recommendation systems, a significant hurdle in the age of big data. Traditional imputation methods struggle to capture complex relationships within the data. We propose a novel approach that uses fine-tuned Large Language Model (LLM) to impute missing values for recommendation tasks. LLM, trained on vast amounts of text, is able to understand complex relationships among data and intelligently fill in missing information. We evaluate our LLM-based imputation method across various tasks within the recommendation system domain, including binary classification, multi-classification, and regression compared to classical data imputation methods. By demonstrating the superiority of LLM imputation over traditional methods, we establish its potential for improving recommendation system performance.

Keywords: Large Language Model, Data Imputation, Recommender System

1 Introduction

The exponential growth of big data has revolutionized many fields, offering unprecedented access to vast amounts of information. Researchers can find tons of information for uncovering patterns and making informed decisions [Belyaeva et al. 2024; Yao et al. 2024]. However, this abundance often masks a hidden adversary: sparse and small data. Missing information, often due to user inactivity, limited data collection, or technical constraints, can significantly hinder the effectiveness of big data models [Fazlikhani et al. 2018]. This is particularly true in recommendation systems, where personalized experiences hinge on a rich understanding of users and items, entries with missing values significantly hinders the ability to generate accurate suggestions [Acharya et al. 2023]. Traditional statistical methods for data imputation, like mean or median imputation, often fall short in capturing the complex relationships and underlying context within the data [Jin et al. 2024a,b].

This paper tackles this challenge by proposing a novel approach that leverages the transformative power of LLM to address the challenge of data imputation in recommendation systems. LLMs, with their remarkable ability to process and understand vast amounts of natural language, possess the potential to intelligently fill in these missing data points. By harnessing the LLM's capability to learn intricate relationships and context from large text corpora, our proposed method aims to impute data that is not only statistically sound but also semantically meaningful [Jäger et al. 2021]. This enriched data can then be utilized by recommendation systems to generate more accurate and personalized suggestions for users.

Focusing on the domain of recommendation systems, we explore the specific application of LLM-based data imputation. Recommender systems rely heavily on comprehensive user and item data to generate personalized suggestions that resonate with individual preferences. By effectively imputing missing values, we aim to create a more complete picture of user behavior and item characteristics. This, in turn, allows the recommendation system to generate more accurate and relevant suggestions, ultimately enhancing the user experience.

We meticulously design a series of experiments to evaluate the effectiveness of our approach. The experiments encompass a diverse range of classification and regression tasks. These experiments delve into binary classification, where the system predicts a single category for an item, multi-classification, which allows for assigning multiple categories, and regression, where the focus is on predicting continuous values like ratings. By demonstrating the superiority of LLM imputation over traditional statistical methods

across these varied scenarios, we aim to establish its significance as a game-changer in improving the performance of recommendation systems.

To comprehensively assess the effectiveness of LLM-based imputation, we conduct rigorous experiments across a diverse range of tasks within the recommender system domain. These experiments encompass binary classification, where the system predicts a single category for an item (e.g., AD recommendation), multi-classification, where multiple categories can be assigned (e.g., multiple categorical movies recommendation), and regression, which focuses on predicting continuous values like ratings or purchase likelihood (e.g., movie rating prediction). By demonstrating the advantage of LLM data imputation over traditional statistical methods in these varied scenarios, we experiment our proposed approach in different recommendations system tasks with different datasets. In summary, our paper makes the following primary contributions:

- We propose a novel approach that utilize LLM to impute missing data which aims to handle data sparsity and data bias issue.
- We further utilize the imputed data and evaluate in the recommendation system which shows improvement to other statistical data imputation strategy.
- Extensive experiment are done to further prove that LLM-based data imputation works better in binary classification, multi-class classification task and regression recommendation tasks.

2 Related Work

2.1 Data Imputation

Missing data presents a pervasive and complex challenge across diverse fields, introducing an additional layer of uncertainty in modeling endeavors. Addressing this issue in modeling tasks can be approached through two primary strategies: employing specialized modeling techniques that directly account for missing data, or utilizing data imputation methods to fill in the gaps [Dempster et al. 1977; Efron 1994; Tian and Porter 2022, 2024; Van Buuren and Groothuis-Oudshoorn 2011]. Data imputation has been studied extensively in both statistics and machine learning, with a rich history of methodological development. Traditional methods, such as replacing missing values with constants (e.g., zero, minimum, maximum) or aggregated measures (mean, median, most frequent), are simple but often introduce bias into the dataset [Newman 2014]. To mitigate this limitation, more sophisticated techniques have been developed, including k-Nearest Neighbors (KNN) imputation, which imputes missing values based on similar data points, and model-based methods that leverage statistical models to predict missing values [Peng and Leng 2024; Sanjar et al. 2020]. Recently, research has focused more on machine learning algorithms for imputation, such as matrix factorization and deep learning techniques [Hwang et al. 2018] , which can handle complex patterns and relationships within the data for more accurate imputations. However, choosing the optimal imputation method remains a critical task, which is influenced by factors such as data type, missing data mechanism (missing completely at random, missing at random, or missing not at random), the amount and pattern of missing data, and specific analytical goals [Ben et al. 2023].

2.2 Large Language Model

LLM trained on massive amounts of text data, have shown promise ability [He et al. 2024] due to their ability to capture complex relationships and semantic information within data. This capability allows them to potentially impute missing values in a more reliable way than traditional methods. For instance, some approaches treat imputation as a classification task, where the LLM predicts the most likely value for the missing entry based on the surrounding data [Li et al. 2024a,b]. Others leverage the generative nature of LLMs to create a distribution of possible values, providing a more comprehensive picture of the imputation uncertainty. While promising, research on LLM-based data imputation is still evolving. Some works showed that LLM has ability to generated data [Zhao et al. 2024]. Areas of exploration include mitigating potential biases present in training data and ensuring the imputed values maintain data integrity, particularly in sensitive domains like healthcare [Deng et al. 2024; Wu et al. 2024]. Overall, LLMs offer a new avenue for tackling missing data issues, with the potential to improve the accuracy and robustness of data analysis in various fields. There are also many successful LLM applications such as in Relation Extration(RE) [Wan et al. 2023], NER [Wang et al. 2023; Xie et al. 2023], feature engineering [Wang et al. 2024a], text summarization [Goyal et al. 2023] and sentiment analysis [Sun et al. 2023].

2.3 Recommender System

Recommender System(RS) is used to generate meaningful suggestions to a collection of users for items or products that might interest them. RS can be divided into personalized [Wu et al. 2022; Yan et al. 2024; Zhao et al. 2022; Zheng et al. 2022] and group-based [Kumar et al. 2022; Sato 2022; Stratigi et al. 2022; Zan et al. 2021; Zhang et al. 2022] systems. In recent years, neural networks like CNN [An and Moon 2022], GCN [Kipf and Welling 2016], GraphSAGE [Hamilton et al. 2017], and others have significantly enhanced RS models [Wang et al. 2024b].

Data sparsity is a persistent challenge in such systems, significantly impacting the accuracy and effectiveness of recommendations. Collaborative filtering techniques, a mainstay in recommendation systems, struggle when user-item interaction matrices are highly sparse, with many missing entries. This sparsity makes it difficult to identify similar users or items for accurate recommendations [Lubos et al. 2024].

More and more research has explored various approaches to address this issue, focus on developing robust recommendation systems that can effectively handle data sparsity and deliver personalized recommendations even with limited user-item interactions. In this paper, we aim to handle those missing data using LLM-based data imputation technology.

3 Method

In this section, we present the major components of our proposed architecture. We start with fine-tuning a pre-trained model using our task-specific dataset that only contains entries without missing values. This fine-tuned model is then employed to impute missing data. The resulting dataset, which contains both the complete and imputed data, is then fed into the recommender system. Figure 1 provides a visual representation of the architecture. Detailed discussions are provided in the subsequent sections.

3.1 Data Preparation

To tailor LLM for our specific task and data at hand, we first need to fine-tune LLM. By fine-tuning a model on a much smaller dataset, its performance on the task can be improved while preserving its general language knowledge. We divide our dataset into two subsets, one is the set that only contains entries without missing values, and the other contains data entries with missing values.

3.2 Fine-tune LLM Model

For the fine-tuning process, we utilize the entries without missing values to enable the model to learn task-specific information. We adopt Low-Rank Adaptation (LoRA) technique [Hu et al. 2022] to achieve efficient fine-tuning of LLM. LLM is typically trained with billions of parameters, rendering comprehensive fine-tuning computationally expensive. LoRA offers a cost-effective alternative by freezing the pre-trained model weights and introducing a set of trainable low-rank adapter parameters. This approach significantly reduces the computational burden associated with fine-tuning while enabling the LLM to adapt to the specific task or domain [Borisov et al. 2023].

The data flow in the fine-tuning process begins with the collection and pre-processing of task-specific data, which is tokenized and converted into input tensors compatible with the LLM architecture. These tensors are then fed into the pre-trained LLM model. Instead of updating the entire weight matrices, LoRA introduces low-rank matrices that approximate the necessary updates. During each forward pass, the input data propagates through the attention and feed-forward layers, where the low-rank matrices are applied to modify the output dynamically. The resulting predictions are compared with the ground truth to compute the loss, which is then backpropagated through the model. Only the parameters associated with the low-rank matrices are updated, leaving the original pre-trained weights largely intact. This selective adaptation allows the model to learn task-specific features efficiently while preserving its general language understanding capabilities learned during LLM's original training. The low-rank matrices focus on the most influential components of the dataset, enhancing the model's ability to predict and fill in missing values accurately. This approach not only speeds up the fine-tuning process but also reduces memory and storage requirements, improving LLM's accuracy on data imputation tasks.

By fine-tuning the pre-trained model with a dataset containing only complete entries, we obtain a LLM that not only retains knowledge from its extensive pre-training but incorporates specific patterns from the current dataset. This approach leverages the model's broad understanding while adapting it to the nuances of our specific task.

3.3 Data Imputation

Subsequently, the fine-tuned LLM mentioned above is used to impute missing data. We incorporate existing data information as relevant knowledge into the prompt. Prompts constructed in this manner contain example-specific information and LLM is used to model the distribution of the missing attributes. Note that the prompt can also be constructed to impute multiple values for a single example simultaneously.For instance, given a data entry with attributes *UserId=11*, *MovieId=44*, *Genres=Sci-FI*, and *Rating=NaN* (indicating missing value), the prompt would be formulated as: *"given a UserID of 11, a MovieID of 44, and a Genre of Sci-Fi, what is the corresponding Rating?"*. As a result, LLM will generate the most probable values based on patterns learnt from the training data and the given prompt. Then *NaN*s are replaced with LLM imputed values. The imputed data is combined with the entries without missing values to form a whole dataset used for training the Recommender System.

3.4 Evaluation in Recommender System

To comprehensively assess the efficacy of the LLM-based data imputation approach, the newly constructed dataset was subsequently employed to train a deep-learning-based recommendation system. To achieve a holistic evaluation of the advantages offered by LLM-based imputation, performance metrics was utilized across various task categories, encompassing binary classification, multi-class classification, and regression. Within the binary classification domain, precision, recall, and F1-score were adopted as the evaluation metrics. For multi-class classification tasks, Recall at k (denoted as R@k) and Normalized Discounted Cumulative Gain at k (denoted as N@k) were employed. Finally, Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) were leveraged to assess the performance of the regression task.

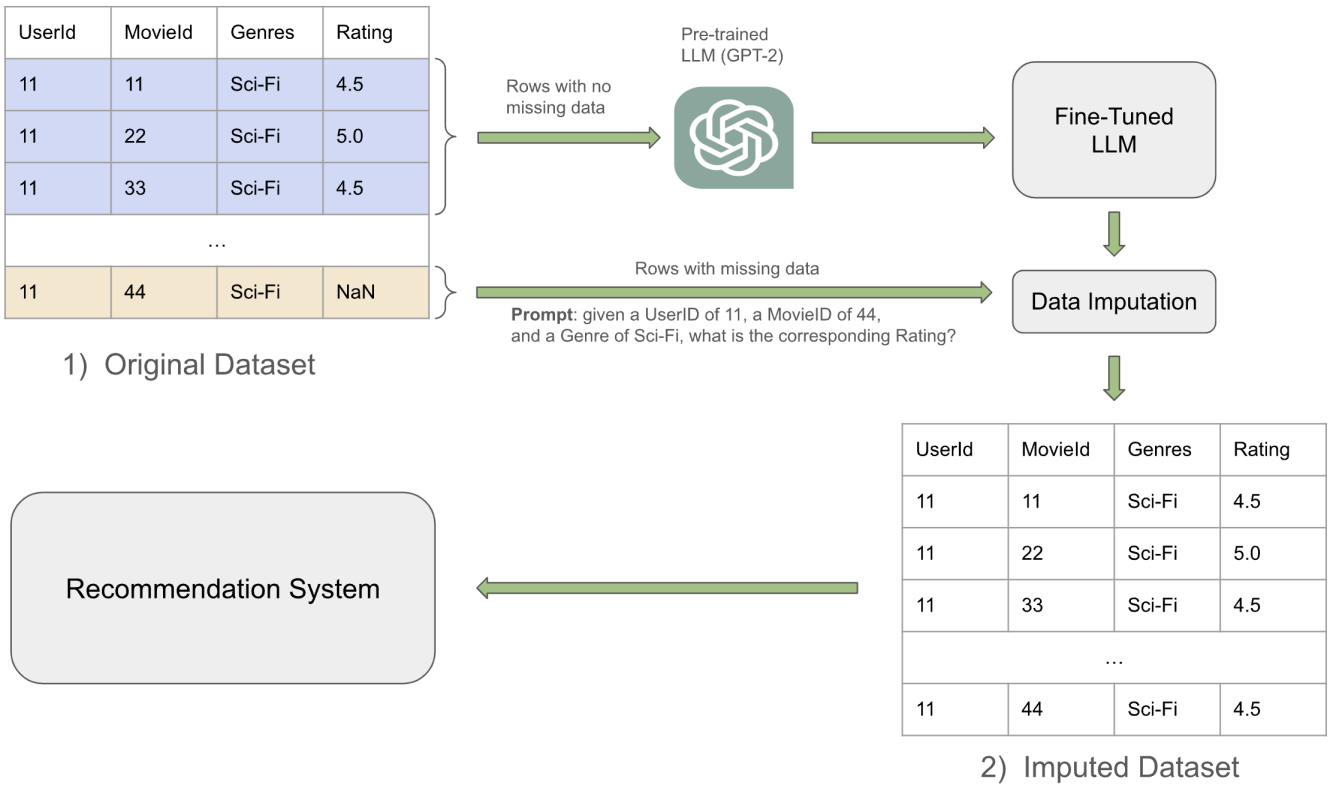

Figure 1. Framework of our proposed method. Original dataset contain missing data. Using entries without missing values to fine-tune LLM which can be further utilized to impute the missing data. After that, complete tabular data are used to feed into Recommender System.

model	R@3↑	R@5↑	R@10↑	N@3↑	N@5↑	N@10↑
Case-Wise Deletion	0.2510	0.3470	0.6050	0.4853	0.5381	0.7011
Zero	0.2370	0.3250	0.5760	0.4412	0.5005	0.6629
Mean	0.2610	0.3490	0.6110	0.5064	0.5455	0.7294
KNN	0.2880	0.3870	0.6420	0.5213	0.5674	0.7331
Multivariate	0.2760	0.3680	0.6440	0.5154	0.5401	0.7420
LLM	**0.2930**	**0.4050**	**0.6530**	**0.5692**	**0.6216**	**0.7632**

Table 1. Comparison among LLM-based and statistical data imputation on multi-class classification task

4 Experiment

4.1 Model and Dataset

We chose to utilize the pre-trained distilled version of GPT-2 model [Sanh et al. 2019] as our LLM because of its open-source accessibility and proven effectiveness across a wide range of tasks. For the choice of dataset, we use AdClick [Jean-Baptiste Tien 2014] and MovieLen [Nacho 2022] dataset for this experiment due to the fact that these are large well-structured dataset without requiring extensive data cleaning. In addition, there have been many researches conducted on these two datasets and they are available for both classification and regression tasks.

The original dataset is a structured tabular dataset. To simulate a dataset with missing values, we introduce missing data in a controlled manner. For each column in the dataset, we randomly select 5% of the data points and mark them as missing. This selection is done independently for each column so that the rows with missing data will vary from column to column. Due to the independent selection process, more than 5% of the rows may contain at least one missing value. Here we break down into 3 different recommendation tasks:

- **Binary Classification:** We leverage the AD Click dataset to evaluate the effectiveness of our proposed

architecture. The imputed data is then fed into a recommendation system designed to classify user clicks on advertisements. This approach aims to improve the accuracy of predicting user engagement with targeted advertising.

- **Multi-class Classification:** We employ the widely used MovieLens dataset to assess the impact of LLM-based data imputation on movie recommendations. The imputed data is subsequently utilized by a recommendation system to suggest a personalized list of top-k movies for each user. This research aims to enhance the effectiveness of recommendation systems by addressing data sparsity issues.
- **Regression:** Building upon the MovieLens dataset, we investigate the use of LLMs for data imputation in predicting user ratings. The imputed data is then incorporated into a recommendation system tasked with predicting user ratings on a scale of 0.0 to 5.0. This approach seeks to improve the accuracy of rating predictions within recommendation systems.

4.2 Baselines

The pre-processed datasets above contain about 5% rows with missing data. To evaluate the effectiveness and efficacy of the proposed LLM-based data imputation, we compare its performance against the following competing baseline methods:

- **Case-Wise Deletion:** without imputing missing data, feed data directly to the recommender system and discard examples with one or more missing values.
- **Zero Imputation:** replaces all missing numeric values with 0.
- **Mean Imputation:** calculates the arithmetic mean of the column and replaces missing values with it.
- **KNN Imputation:** imputes missing values using k-Nearest Neighbors. Each sample's missing values are imputed using the mean value from n-th nearest neighbors found in the training set.
- **Multivariate Imputation:** estimates each feature from all the others and imputes missing values by modeling each feature with missing values as a function of other features in a round-robin fashion [Pedregosa et al. 2011]. Multiple imputation is known to preserve the relationship among variables. Also, the uncertainty associated with the missing values is considered in the data imputation process[Van Buuren and Oudshoorn 2000].

4.3 Evaluation

To better evaluate LLM-based data imputation technique, we evaluate our model's performance across three tasks: binary classification, multi-class classification, and regression. Two benchmark datasets, AD click and MovieLens, are used. For both datasets, we meticulously curate the data to achieve a targeted missing value ratio of approximately 5%. Then, we feed the data with no imputation into our recommendation system for baseline performance. The DLRM [Naumov et al. 2019] model is utilized for this purpose and we randomly split data with 60/20/20 ratio for training, testing, and validation, respectively. In addition, we applied statistical methods (mean, zero, KNN, and Multivariate) and our LLM-based approaches. The imputed data by different approaches will feed into DLRM one by one following the same 60/20/20 ratio for the evaluation.

model	precision↑	recall↑	f1-score↑
Case-Wise Deletion	0.1980	0.4450	0.2740
Zero	0.7207	0.7200	0.7200
Mean	0.8846	0.8700	0.8702
KNN	**0.9192**	**0.9150**	**0.9150**
Multivariate	0.8970	0.8900	0.8903
LLM	<u>0.9071</u>	<u>0.9001</u>	<u>0.9003</u>

Table 2. Comparison among LLM-based and statistical data imputation on binary classification task

The detailed results of binary classification task are presented in Table 2, with the top and second-highest performing models highlighted for clarity. In the context of binary classification, models constructed using KNN imputation demonstrated superior performance, while our proposed method achieved the second-best results. This outcome may be attributed to the dichotomous nature of the problem, which potentially favors approaches that identify similar instances. However, it is important to note that such methods may not maintain their efficacy in more complex, multi-class scenarios. Table 1 presents the results of multi-class classification task. Due to the richer metadata and intricate relationships within the MovieLens dataset, the LLM-based model demonstrates a clear advantage over other models. Finally, we evaluate the effectiveness of LLM-based data imputation within a regression task comparing with statistical methods. Table 3 showcases the results, highlighting the superior performance of the LLM-based data imputation approach compared to other models.

Finally, we evaluate the effectiveness of LLM-based data imputation within a regression task comparing with statistical methods. Table 3 showcases the results, highlighting the superior performance of the LLM-based data imputation approach compared to other models.

5 Conclusion

In conclusion, this paper proposes a novel approach that leverages the power of LLM to address missing data in the Recommender System. By imputing missing data in a semantically meaningful way, our method enriches data and allows the Recommender System to generate more accurate and

model	MAE↓	MSE↓	RMSE↓
Case-Wise Deletion	0.7659	0.9792	0.9895
Zero	0.7798	0.9928	0.9964
Mean	0.7804	0.9883	0.9942
KNN	0.7791	0.9909	0.9955
Multivariate	0.7785	0.9887	0.9943
LLM	**0.7612**	**0.9647**	**0.9822**

Table 3. Comparison among LLM-based and statistical data imputation on regression task

personalized suggestions, ultimately enhancing user experience. We extensively evaluate our approach across various recommender system tasks, demonstrating its effectiveness in improving performance compared to traditional data imputation methods. The implications of this research extend beyond recommender systems, opening new avenues for utilizing LLMs to mitigate data sparsity and small sample size issues in big data models, leading to a more robust Recommender System.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. LLM Based Generation of Item-Description for Recommendation System. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) *(RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1204–1207. https://doi.org/10.1145/3604915.3610647

Hyeon-woo An and Nammee Moon. 2022. Design of recommendation system for tourist spot using sentiment analysis based on CNN-LSTM. *Journal of Ambient Intelligence and Humanized Computing* 13, 3 (2022), 1653–1663.

Anastasiya Belyaeva, Justin Cosentino, Farhad Hormozdiari, Krish Eswaran, Shravya Shetty, Greg Corrado, Andrew Carroll, Cory Y. McLean, and Nicholas A. Furlotte. 2024. Multimodal LLMs for Health Grounded in Individual-Specific Data. In *Machine Learning for Multimodal Healthcare Data*, Andreas K. Maier, Julia A. Schnabel, Pallavi Tiwari, and Oliver Stegle (Eds.). Springer Nature Switzerland, Cham, 86–102.

Ângela Jornada Ben, Johanna M. van Dongen, Mohamed El Alili, Martijn W. Heymans, Jos W. R. Twisk, Janet L. MacNeil-Vroomen, Maartje de Wit, Susan E. M. van Dijk, Teddy Oosterhuis, and Judith E. Bosmans. 2023. The handling of missing data in trial-based economic evaluations: should data be multiply imputed prior to longitudinal linear mixed-model analyses? *The European Journal of Health Economics* 24, 6 (01 Aug 2023), 951–965. https://doi.org/10.1007/s10198-022-01525-y

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023. Language Models are Realistic Tabular Data Generators. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=cEygmQNOeI

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: series B (methodological)* 39, 1 (1977), 1–22.

Qixin Deng, Qikai Yang, Ruibin Yuan, Yipeng Huang, Yi Wang, Xubo Liu, Zeyue Tian, Jiahao Pan, Ge Zhang, Hanfeng Lin, et al. 2024. ComposerX: Multi-Agent Symbolic Music Composition with LLMs. *arXiv preprint arXiv:2404.18081* (2024).

Bradley Efron. 1994. Missing data, imputation, and the bootstrap. *J. Amer. Statist. Assoc.* 89, 426 (1994), 463–475.

Fatemeh Fazlikhani, Pegah Motakefi, and Mir Mohsen Pedram. 2018. Missing Data Imputation by LOLIMOT and FSVM/FSVR Algorithms with a Novel Approach: A Comparative Study. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Foundations*, Jesús Medina, Manuel Ojeda-Aciego, José Luis Verdegay, David A. Pelta, Inma P. Cabrera, Bernadette Bouchon-Meunier, and Ronald R. Yager (Eds.). Springer International Publishing, Cham, 551–569.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News Summarization and Evaluation in the Era of GPT-3. arXiv:2209.12356 [cs.CL]

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems* 30 (2017).

Nan He, Hanyu Lai, Chenyang Zhao, Zirui Cheng, Junting Pan, Ruoyu Qin, Ruofan Lu, Rui Lu, Yunchen Zhang, Gangming Zhao, Zhaohui Hou, Zhiyuan Huang, Shaoqing Lu, Ding Liang, and Mingjie Zhan. 2024. TeacherLM: Teaching to Fish Rather Than Giving the Fish, Language Modeling Likewise. arXiv:2310.19019 [cs.CL] https://arxiv.org/abs/2310.19019

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

Won Seok Hwang, Shu Li, Seung Woo Kim, and Keun Ho Lee. 2018. Data imputation using a trust network for recommendation via matrix factorization. *Computer Science and Information Systems* 15, 2 (2018), 347–368.

Olivier Chapelle Jean-Baptiste Tien, joycenv. 2014. Display Advertising Challenge. https://kaggle.com/competitions/criteo-display-ad-challenge

Can Jin, Tong Che, Hongwu Peng, Yiyuan Li, and Marco Pavone. 2024a. Learning from teaching regularization: Generalizable correlations should be easy to imitate. *arXiv preprint arXiv:2402.02769* (2024).

Can Jin, Hongwu Peng, Shiyu Zhao, Zhenting Wang, Wujiang Xu, Ligong Han, Jiahui Zhao, Kai Zhong, Sanguthevar Rajasekaran, and Dimitris N Metaxas. 2024b. APEER: Automatic Prompt Engineering Enhances Large Language Model Reranking. *arXiv preprint arXiv:2406.14449* (2024).

Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. 2021. A Benchmark for Data Imputation Methods. *Frontiers in Big Data* 4 (2021). https://doi.org/10.3389/fdata.2021.693674

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

Chintoo Kumar, C Ravindranath Chowdary, and Deepika Shukla. 2022. Automatically detecting groups using locality-sensitive hashing in group recommendations. *Information Sciences* 601 (2022), 207–223.

Xinjin Li, Jinghao Chang, Tiexin Li, Wenhan Fan, Yu Ma, and Haowei Ni. 2024a. A Vehicle Classification Method Based on Machine Learning. *Preprints* (July 2024). https://doi.org/10.20944/preprints202407.0981.v1

Xinjin Li, Yuanzhe Yang, Yixiao Yuan, Haowei Ni, Yu Ma, and Yangchen Huang. 2024b. Intelligent Vehicle Classification System Based on Deep Learning and Multi-Sensor Fusion. *Preprints* (July 2024). https://doi.org/10.20944/preprints202407.2102.v1

Sebastian Lubos, Thi Ngoc Trang Tran, Alexander Felfernig, Seda Polat Erdeniz, and Viet-Man Le. 2024. LLM-generated Explanations for Recommender Systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 276–285. https://doi.org/10.1145/3631700.3665185

Nacho. 2022. Movie recommender system (2022). https://kaggle.com/competitions/movie-recommender-system-2022

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

Daniel A Newman. 2014. Missing data: Five practical guidelines. *Organizational Research Methods* 17, 4 (2014), 372–411.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Lai Peng and Qian Leng. 2024. Research on the Application of Support Vector Machine Algorithm Model With Multi-Modal Data Fusion in Breast Cancer Ultrasound Image Classification. *Applied and Computational Engineering* 67, 1. https://doi.org/10.54254/2755-2721/67/20240671

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*.

Karshiev Sanjar, Olimov Bekhzod, Jaesoo Kim, Anand Paul, and Jeonghong Kim. 2020. Missing Data Imputation for Geolocation-based Price Prediction Using KNN–MCF Method. *ISPRS International Journal of Geo-Information* 9, 4 (2020). https://doi.org/10.3390/ijgi9040227

Ryoma Sato. 2022. Enumerating fair packages for group recommendations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 870–878.

Maria Stratigi, Evaggelia Pitoura, Jyrki Nummenmaa, and Kostas Stefanidis. 2022. Sequential group recommendations based on satisfaction and disagreement scores. *Journal of Intelligent Information Systems* (2022), 1–28.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment Analysis through LLM Negotiations. arXiv:2311.01876 [cs.CL]

Jiahao Tian and Michael D Porter. 2022. Changing presidential approval: Detecting and understanding change points in interval censored polling data. *Stat* 11, 1 (2022), e463.

Jiahao Tian and Michael D Porter. 2024. Time of week intensity estimation from partly interval censored data with applications to police patrol planning. *Journal of Applied Statistics* (2024), 1–19.

Stef Van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 45 (2011), 1–67.

Stef Van Buuren and Catharina GM Oudshoorn. 2000. Multivariate imputation by chained equations.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3534–3547. https://doi.org/10.18653/v1/2023.emnlp-main.214

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named Entity Recognition via Large Language Models. arXiv:2304.10428 [cs.CL]

Yining Wang, Jinman Zhao, and Yuri Lawryshyn. 2024a. GPT-Signal: Generative AI for Semi-automated Feature Engineering in the Alpha Research Process. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*. Jeju, South Korea, 42–53. https://aclanthology.org/2024.finnlp-2.4

Zeyu Wang, Yue Zhu, Zichao Li, Zhuoyue Wang, Hao Qin, and Xinqi Liu. 2024b. Graph Neural Network Recommendation System for Football Formation. *Applied Science and Biotechnology Journal for Advanced Research* 3, 3 (May 2024), 33–39. https://doi.org/10.5281/zenodo.12198843

Chao Wu, Sannyuya Liu, Zeyu Zeng, Mao Chen, Adi Alhudhaif, Xiangyang Tang, Fayadh Alenezi, Norah Alnaim, and Xicheng Peng. 2022. Knowledge graph-based multi-context-aware recommendation algorithm. *Information Sciences* 595 (2022), 179–194.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-Any Multimodal LLM. arXiv:2309.05519 [cs.AI] https://arxiv.org/abs/2309.05519

Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, and Hongwei Wang. 2023. Empirical Study of Zero-Shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 7935–7956. https://doi.org/10.18653/v1/2023.emnlp-main.493

Yubing Yan, Camille Moreau, Zhuoyue Wang, Wenhan Fan, and Chengqian Fu. 2024. Transforming Movie Recommendations with Advanced Machine Learning: A Study of NMF, SVD,and K-Means Clustering. arXiv:2407.08916 [cs.LG] https://arxiv.org/abs/2407.08916

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4, 2 (2024), 100211. https://doi.org/10.1016/j.hcc.2024.100211

Shuxun Zan, Yujie Zhang, Xiangwu Meng, Pengtao Lv, and Yulu Du. 2021. UDA: A user-difference attention for group recommendation. *Information Sciences* 571 (2021), 401–417.

Song Zhang, Nan Zheng, and Danli Wang. 2022. GBERT: Pre-training user representations for ephemeral group recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2631–2639.

Chenyang Zhao, Xueying Jia, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. SELF-GUIDE: Better Task-Specific Instruction Following via Self-Synthetic Finetuning. arXiv:2407.12874 [cs.CL] https://arxiv.org/abs/2407.12874

Rongmei Zhao, Shenggen Ju, Jian Peng, Ning Yang, Fanli Yan, and Siyu Sun. 2022. Two-level graph path reasoning for conversational recommendation with user realistic preference. In *proceedings of the 31st ACM international conference on information & knowledge management*. 2701–2710.

Jiayin Zheng, Juanyun Mai, and Yanlong Wen. 2022. Explainable session-based recommendation with meta-path guided instances and self-attention mechanism. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2555–2559.