User Preferences for Large Language Model versus Template-Based Explanations of Movie Recommendations: A Pilot Study

Julien Albert
UNamur
julien.albert@unamur.be

Martin Balfroid
UNamur
martin.balfroid@unamur.be

Miriam Doh
ULB-UMONS
miriam.doh@ulb.be

Jeremie Bogaert
UCLouvain
jeremie.bogaert@uclouvain.be

Luca La Fisca
UMONS
luca.lafisca@umons.be

Liesbet De Vos
UNamur
liesbet.devos@unamur.be

Bryan Renard
Multitel-UNamur
renard@multitel.be
bryan.renard@unamur.be

Vincent Stragier
UMONS
vincent.stragier@umons.ac.be

Emmanuel Jean
Multitel
jean@multitel.be

***Abstract*—We daily interact with recommender systems, from online shopping to streaming services, yet we often question the reason behind certain recommendations. While these systems may offer textual explanations to clarify their recommendations, the rule-based approach often fails to satisfy the user. Our pilot study, involving 25 participants, addresses this gap by comparing the traditional template-based approach to a more dynamic method that employs a large language model (LLM) for explanation generation. Additionally, the study explores variations in the LLM approach, such as rephrasing a provided template or using a knowledge graph for context. Although subject to high variance, preliminary findings suggest that LLM-generated explanations may offer a more nuanced and engaging user experience, better aligning with user expectations. This study sheds light on the potential limitations of current explanation methods and offers promising directions for leveraging large language models to improve user satisfaction and trust in recommender systems.**

***Index Terms*—Large Language Models, Recommender Systems, Explainability, GD6**

I. Introduction

Most of us wonder daily why platforms like Facebook and YouTube recommend specific people or videos to us. The lack of transparency in these recommendations often leaves us without a clear explanation. This can degrade user confidence, recommendation acceptance and, more broadly, the user experience [1]. To address those important concerns, a growing field of research focuses on making recommendation systems more transparent and explainable [1]–[3]. A promising approach is to use large language models (LLMs) to generate explanations for recommendations. LLMs are initially pre-trained on extensive corpora, allowing them to perform a versatile range of natural language processing (NLP) tasks [4]. The generated text is typically well-written and clear, making it easy for humans to understand.

Motivated by these perspectives, we put them to the test in the generation of explanations for recommendations during the TRAIL'23 Workshop[1]. Concretely, we defined two goals to address during the workshop. The first is to implement working examples of recommendation explanations generated with LLMs using various recommendation methods and LLM models. This way, we could assess the technical possibilities and limitations of LLMs. The second goal is to evaluate explanations generated by different LLM models and recommendation methods to understand their qualities and their limitations in this context. To achieve this goal, we designed a user-based evaluation method to assess explanations w.r.t. different explanatory goals and subjective properties [1].

II. Technical Exploration and Implementation

As shown in Fig. 1, we propose a pipeline that takes user preferences (i.e., past interactions with items) as input, and generates explained recommendations as output. The most important design choice is to separate the recommendation and explanation processes, only using LLMs to explain items previously recommended by an independant recommendation method. We choose to use classic recommendation to ensure valid recommendation, as hallucination is an important issue with LLMs [5]. Moreover, this choice allows us to isolate the explanation task, empowering us to compare explanations created by a baseline explanation method, with explanations written by LLMs.

Regarding the recommendation methods, we focused on graph-based methods. More specifically, we used *Personalized PageRank* [6] and RippleNet [7], both of which generate explanations based on a graph of the past interactions between users and items. This graph is augmented with knowledge about the movie domain, to further guide the recommendation system. Those methods also provide explanations for recommendations in the form of paths from the seed items to the recommended

[1]https://trail.ac/en/trail-summer-workshops/the-trail-summer-workshop-2023/, more details in the Appendix

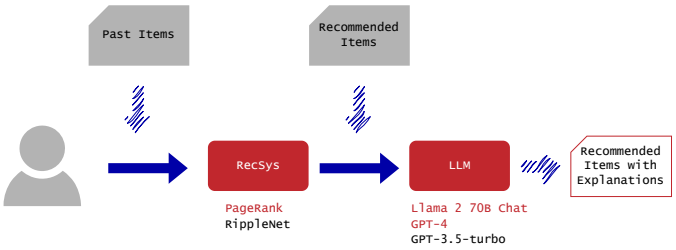

Fig. 1. Pipeline used to guide our experiments. Methods and models used for the evaluation part are in fuchsia.

ones. The datasets used for experimentation were Movielens-1M[2] and MindReader[3]. For the user-based evaluation, we only used Movielens-1M in combination with the Personalized PageRank.

We are interested in textual explanations, since they convey rich information to the user [1]. The main existing approaches are template-based and generation-based [3]. As a baseline, we use a template-based approach, which transcripts path-based explanations into text. We compare this baseline method to LLM-based methods for generating explanations, inspired by the literature on the topic, e.g., PEPLER [8].

Large Language Models (LLMs) are now some of the world's most famous NLP models due to the publicity made by OpenAI with ChatGPT, which uses LLMs, i.e., GPT-3.5-turbo and GPT-4 (SOTA). They can perform various NLP tasks. Current LLMs use a decoder-only architecture based on the transformer's architecture [9]. They are trained to give a probability of distribution over the vocabulary of tokens, allowing to predict the next token. The tokens are subparts of sentences, and the vocabulary of tokens, fixed and based on the training data, is often built using byte pair encoding (BPE) [10], [11]. To produce sequences of tokens, we used greedy decoding with Llama 2 70B Chat and the default technique (which we don't know of) when using GPT-4. Greedy decoding only considers the most probable token at each generation step, which is time-efficient, unlike other techniques. We decided on using greedy decoding due to time constraints we had during the TRAIL'23 summer workshop.

We considered two methods for generating explanations for movie recommendations. We aimed to measure how effectively each approach could deliver concise yet informative explanations to users that align with their expectations.

Three types of explanations were finally kept for the user-based evaluation (as shown in Fig. 2):

1) **Template-based**: our baseline method, which uses a template to generate explanations algorithmically based on the edges and nodes of the explanation paths;
2) **LLM-based**: which uses LLMs to generate the explanation. We explored two variations:
   a) **LLM-based rephrasing**: rephrase the template-based explanation;

[2]https://grouplens.org/datasets/movielens/
[3]https://mindreader.tech/dataset/

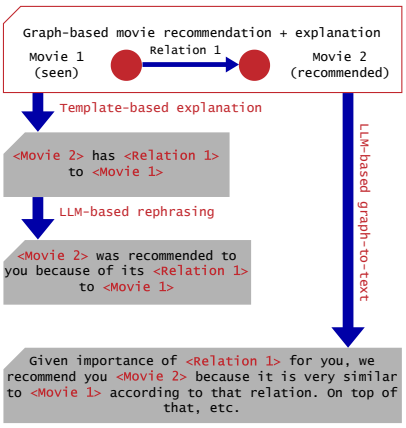

Fig. 2. Illustration of the three types of explanations compared in the user evaluation.

   b) **LLM-based graph-to-text**: the model deduces the reasoning behind the recommendation given a knowledge graph as context.

Between the two LLM variants, only the context varies, either the template-based explanation or the graph. The definition of the task is, therefore, the same for both: to explain why a particular film has been recommended. To ensure a fairly consistent format across each generation, we constrained the LLM's behaviour [12] by specifying that only one paragraph should be used and that it should be written in layman's terms. Otherwise, the model tended to ramble and use technical terms that could confuse the user.

```
"""We recommended "The Hunger Games: Mockingjay - Part 1" because:
"The Hunger Games: Mockingjay - Part 1" is from decade Movies of the 2010s
like "A Quiet Place"""

(Explain in one paragraph and in layman's terms why 'The Hunger Games:
Mockingjay - Part 1' was recommended:)
```

(a) LLM-based rephrasing

```
[['BURN-E', 'FROM_DECADE', 'Decade-2000'], ['A Quiet Place',
'FROM_DECADE', 'Decade-2010'], ['Decade-2000', 'FROM_DECADE', 'Final
Destination 3'], ['Adventure Film', 'HAS_GENRE', 'The Hunger Games'], ['Bolt',
'HAS_GENRE', 'Adventure Film'], ['Bolt', 'FROM_DECADE', 'Decade-2000'], ['The
Final Destination', 'FOLLOWED_BY', 'Final Destination 3'], ['Horror Film',
'HAS_GENRE', 'A Quiet Place'], ['The Hunger Games: Catching Fire',
'FOLLOWED_BY', 'The Hunger Games'], ['Decade-2010', 'FROM_DECADE', 'The Hunger
Games'], ['The Hunger Games: Catching Fire', 'HAS_GENRE', 'Science Fiction
Film'], ['The Final Destination', 'HAS_GENRE', 'Horror Film'], ['The Hunger
Games: Catching Fire', 'FOLLOWED_BY', 'The Hunger Games: Mockingjay - Part
1'], ['Decade-2010', 'FROM_DECADE', 'The Hunger Games: Mockingjay - Part 1'],
['Science Fiction Film', 'HAS_GENRE', 'BURN-E'], ['Adventure Film',
'HAS_GENRE', 'The Hunger Games: Catching Fire']]

(Explain in one paragraph and in layman's terms why 'The Hunger Games:
Mockingjay - Part 1' was recommended:)
```

(b) LLM-based graph-to-text

Fig. 3. Here is an example of the same recommendation presented in the same format as the prompt in Liu et al. [13]. **Black**-colored text outlines the task, **red**-colored text highlights the formatting guidelines, and **blue**-colored text is either the given template or the graph.

III. User-based evaluation

A. Methodology

Part of our project's goal was to perform a user-based evaluation of the three types of explanations generated by our pipeline. We drew inspiration from [14] to craft the structure

for our evaluation procedure (Fig. 4), albeit with slight modifications due to the inclusion of LLM-generated explanations. We decided to focus on the following key aspects:

1) Assessing user expectations of recommendation explanations using the seven goals from [15], also used by [14].
2) Presenting a recommended item to the user alongside multiple alternative explanations (based on a watching profile selected by the user beforehand).
3) Requesting users to assess the explanations based on their general preference and measure the extent to which each explanation satisfies the seven goals.
4) Gathering qualitative insights via open question on user expectations and explanation assessments.

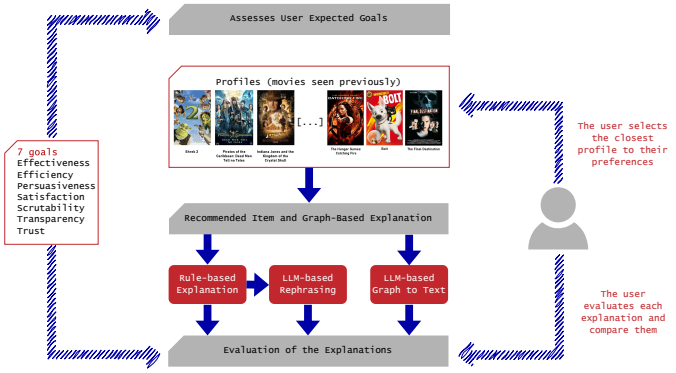

Fig. 4. Structure of our user evaluation procedure

Afterward, we conducted a dry run to validate the clarity of the questionnaire and assess its length to prevent evaluator fatigue. Subsequently, we rolled out the questionnaire to multiple evaluators to gather their responses.

B. Results

We conducted 25 user tests with TRAIL'23 Workshop participants (researchers in AI). The small number of participants means that no statistically robust conclusions can be drawn, but certain trends can be observed.

Concerning the user expectations about explanations, we observe no difference in importance for the seven goals investigated. However, concerning user assessment of the generated explanations (see Fig. 5), we observe that the explanation generated by the LLM from a knowledge graph performs best w.r.t. of the 7 goals. And this result is confirmed by the participants' general assessment of the explanations. According to the participants, this explanation type is mainly preferred because it's often more detailed and more pleasant to read. However, beyond the small sample size ($n = 25$), it is important to point out the significant variance in these last results. This indicates strong differences between participants in the way they perceive explanations, which is a result that should be investigated further.

We also discovered that LLMs often introduce additional, usually accurate, information in their explanations based on movie titles. This is unsurprising given the model's capacity to draw from cultural references [12]. This information may

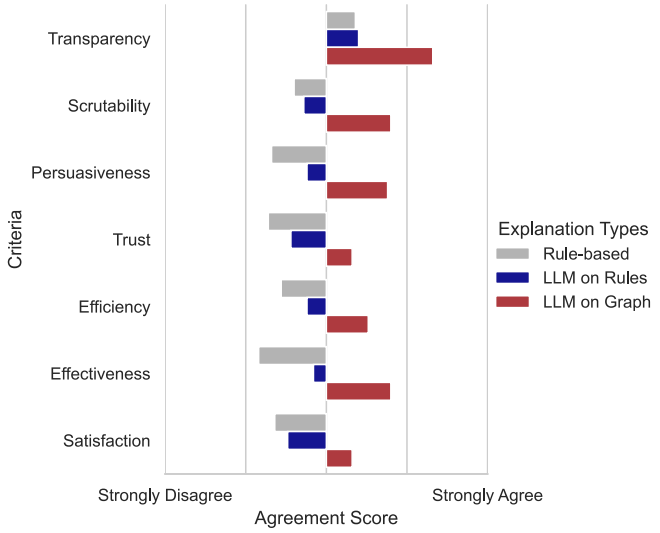

Fig. 5. User assessment of explanations w.r.t. the 7 goals. about the recommendation explanations.

be desirable, depending on whether the user prefers enriched content, or a contrary explanations based on the recommendation system's logic only. Nonetheless, this aspect can easily be controlled by replacing movie titles with placeholder labels when generating explanations that should not contain movie-specific knowledge added by the model itself.

IV. Future Works

In the future, we would like to explore various models. In particular, we would like to focus on smaller models, to understand how model size affects efficiency. We expect that, due to their size, smaller models may struggle to generate explanations based on the knowledge graph, but may be sufficient for rephrasing the template-based explanations. This is particularly relevant considering the substantial computational requirements of larger LLMs. Furthermore, fine-tuning could be explored, as well as advanced prompting techniques like chain-of-thought [16] and self-consistency [17] (which may appeal to users wanting more detailed reasoning). Another interesting approach might be to specify the explanation generation task by memetic proxy [12], i.e., to use the model's ability to draw on cultural references, metaphors, analogies, role-playing, and so on.

Finally, instead of only relying on user-based evaluations, we aim to use mixed-methods evaluation to draw a complete picture of LLM's explanation generation capabilities for recommendations. This evaluation would combine heuristics-based methods (based on classical metrics for text quality like BLEU [18] and ROUGE [19] scores), explanation quality metrics (e.g., [8]), and user-based methods. Such user-based methods could include qualitative (e.g., interviews) and quantitative (e.g., online survey) methods to assess explanations w.r.t. different explanatory goals and subjective properties [1].

APPENDIX

A. Authors' Biographies

1) Julien Albert: After an initial career as a librarian, Julien Albert embarked on a career change and obtained a master's degree in computer science from UNamur in 2020. He then worked for one year at UNamur on the EFFaTA-MeM research project, which aims to develop innovative tools for text analysis. In September 2021, he began a Ph.D. in computer science at UNamur under the supervision of Professors Benoît Frenay and Bruno Dumas. His research area is explainability in artificial intelligence. His approach involves placing the user at the centre of concerns by combining explainability techniques from machine learning with methods developed in human-computer interaction.

2) Martin Balfroid: Martin Balfroid is a PhD student at the University of Namur, his research investigates AI-in-the-loop approaches to improve software engineering. He earned his master's in Computer Science, focusing on Data Science, in June 2022. The results of his master's thesis were published at the 2nd Software Testing Education Workshop. Martin began his PhD in July 2022 with funding from the ARIAC project and is supervised by Assistant Professors Benoît Vanderose and Xavier Devroey.

3) Miriam Doh: Miriam Doh obtained a master's degree in Information and Communication Engineering from the University of Trento (UniTn) in Italy in 2021. Her master's thesis focused on the application of genetic algorithms to social networks, with a focus on studying the problem of community segregation in metropolitan areas. After completing her degree, she began a joint PhD program between ULB and UMONS on the intersection of Deep Learning and Computer Vision, with a particular emphasis on Explainable AI (XAI). Her research project is dedicated to exploring the integration of Cognitive Psychology principles to advance Explainable and Trustworthy Artificial Intelligence, particularly within the context of Face Analysis applications.

4) Jeremie Bogaert: Jérémie Bogaert obtained his master's degree in computer science engineering with a focus on artificial intelligence from UCLouvain in 2021. His master's thesis explored the limitations of deep fake news generation models and their detection using machine learning models and human readers. He started his doctoral thesis at UCLouvain in September 2021 and is currently working on the interaction between interpretable machine learning models and human readers for the detection of deep fake news.

5) Luca La Fisca: Luca La Fisca is currently a PhD student with a keen interest in neural engineering. His primary research focus revolves around advancing tools for a deeper understanding of the intricacies of the human brain. Luca's doctoral thesis specifically delves into the realm of ElectroEncephalogram (EEG) analysis. He is particularly fascinated by the interpretation of latent space to unveil critical interactions among brain regions during the execution of specific tasks, with a primary emphasis on visual tasks. Additionally, Luca harbours a strong interest in the field of neurofeedback.

Within the ARIAC project, Luca La Fisca is actively involved in Work Package 1, which centres on the interactions between humans and artificial intelligence. His contributions span various aspects, including interactive and human-in-the-loop algorithms, user assistance in AI-in-the-loop scenarios, consensus mechanisms, handling imperfect multi-expert labels, and the development of explainable AI solutions.

6) Liesbet De Vos: Liesbet De Vos obtained a master's in Linguistics at the Catholic University of Leuven in 2021. Fascinated by computational linguistics, she completed her studies with an advanced master's in Artificial Intelligence at the Catholic University of Leuven, which she completed in 2022. Liesbet continues to nurture her passion for language during a PhD at the University of Namur, where she focuses on building hybrid AI systems that learn to use language through the same mechanisms as humans. In her thesis, she aims to extend the computational construction grammar framework to the visual modality so that it can adequately represent and learn the linguistic structure of sign languages. Within the ARIAC project, Liesbet actively contributes to Work Package 2, which revolves around trust mechanisms for artificial intelligence.

7) Bryan Renard: Bryan Renard obtained a master's degree in theoretical physics from UNamur in 2022. He then changed his career path and is now a dedicated PhD student whose research interests span several exciting domains within the field of artificial intelligence. His primary focus is the application of artificial intelligence in the realm of proteins, exploring innovative ways to harness AI (especially LLMs) for protein-related research. Additionally, Bryan is passionate about self-supervised learning, particularly in the context of Automatic Speech Recognition (ASR).

His thesis is jointly conducted by UNamur and Multitel. It is funded by the FoodWal portfolio from the Public Service of Wallonia (Economy, Employment, and Research), more particularly within the PEPTIBoost project. As a part of the ARIAC project, Bryan Renard plays an integral role in Work Package 4, which revolves around optimizing AI implementations. His contributions encompass a wide range of topics, including transfer learning, High-Performance Computing (HPC) and self-supervised learning techniques.

8) Vincent Stragier: Vincent Stragier is a PhD student at the University of Mons (UMONS). He is working on an interactive assistant for visually impaired and blind people within the ISIA Lab, a department of the Faculty of Engineering. His research interests are mainly focused on NLP, large language models and computer vision related topics.

In 2021, he obtains his master's degree in electrical Engineering, specialized in Signals, Systems and BioEngineering from the Faculty of Engineering in Mons. In 2020, he works on an epilepsy detection pipeline base on an XGBoost classifier built by the CETIC, where he is Engineer Intern at the time. During his studies, he participates in the electronic student association, electroLAB, and the Erasmus Student Network of Mons, ESNMons. In his free time, he likes taking photographs,

fixing various things (hardware and software related), and learning new skills.

9) Emmanuel Jean: In 2009, Emmanuel Jean earned a dual degree in electrical engineering from the Faculty of Engineering at the University of Mons and Supelec-Paris. Subsequently, he joined the Signal Processing and Embedded Systems department at Multitel, where he actively participated in various regional and European projects involving vocal technologies and multimodal Human-Computer Interaction (HCI).

In 2012, he furthered his education by obtaining a Bachelor's degree in Management Sciences from the Louvain School of Management at UCL-Mons. Since 2017, his professional focus has shifted towards diverse projects centred around Deep Learning applied to temporal signals, including audio, speech, and vibrations. His current research interests revolve around the development of weakly supervised machine learning techniques and the deployment of reliable artificial intelligence systems.

B. ARIAC and TRAIL

TRAIL and the ARIAC research project are part of the regional DigitalWallonia4.ai program, which aims to accelerate the development of artificial intelligence technologies in Wallonia.

TRAIL (TRusted AI Labs) provides actors in the socio-economic landscape with R&D expertise and AI technological bricks developed by the 5 French-speaking universities and 4 approved research centres active in AI. To achieve this, the SPW-EER has allocated a budget of €32 million for the ARIAC research project led by the TRAIL consortium. This initiative is part of the 4th axis of the regional DigitalWallonia4.ai programme: "Research, innovation and partnerships".

The ambition is to pool research in artificial intelligence in the Wallonia-Brussels Federation and is concretely reflected through the research project "ARIAC by DigitalWallonia4.ai", based on an agreement between the Walloon Region (SPW Research) and the actors forming the TRAIL consortium.

The ARIAC project ("Applications and Research for Trusted Artificial Intelligence" in English or "Applications et Recherche pour une Intelligence Artificielle de Confiance" in French) is spread over 6 years and is articulated around 5 WP (Work Package):

- human-AI interaction,
- trust mechanisms for AI,
- model-AI integration,
- optimized implementations of AI,
- TRAIL Factory.

ACKNOWLEDGMENT

This research was partially supported by the ARIAC project (No. 2010235), funded by the Service Public de Wallonie (SPW Recherche). This research used resources of the "Plateforme Technologique de Calcul Intensif (PTCI)" (http://www.ptci.unamur.be) located at the University of Namur, Belgium, which is supported by the FNRS-FRFC, the Walloon Region, and the University of Namur (Conventions No. 2.5020.11, GEQ U.G006.15, 1610468, RW/GEQ2016 et U.G011.22). The PTCI is member of the "Consortium des Équipements de Calcul Intensif (CÉCI)" (https://www.ceci-hpc.be). Vincent Stragier is funded through a PhD grant from the Œuvre fédérale Les Amis des Aveugles et Malvoyants ASBL- The Friends of the Blind and Visually Impaired Federal Charity-, Ghlin, Belgium and the Loterie Nationale, Rue Belliard 25-33, 1040 Brussels, Belgium. Vincent Stragier is partially supported by the FNRS-FRS. Bryan Renard is funded by the Public Service of Wallonia (Economy, Employment and Research), under the FoodWal agreement n°2210182 from the Win4Excellence project of the Wallonia Recovery Plan.

REFERENCES

[1] N. Tintarev and J. Masthoff, "Explaining Recommendations: Design and Evaluation," in *Recommender Systems Handbook*, pp. 353–382, Boston, MA: Springer US, 2015.

[2] A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos, "A generalized taxonomy of explanations styles for traditional and social recommender systems," *Data Mining and Knowledge Discovery*, vol. 24, pp. 555–583, may 2012.

[3] Y. Zhang and X. Chen, "Explainable Recommendation: A Survey and New Perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[4] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv e-prints*, pp. arXiv–2108, 2021.

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, mar 2023.

[6] T. Haveliwala, "Topic-sensitive pagerank: a context-sensitive ranking algorithm for web search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784–796, 2003.

[7] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, and M. Guo, "Ripplenet: Propagating user preferences on the knowledge graph for recommender systems," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, (New York, NY, USA), p. 417–426, Association for Computing Machinery, 2018.

[8] L. Li, Y. Zhang, and L. Chen, "Personalized Prompt Learning for Explainable Recommendation," *ACM Transactions on Information Systems*, vol. 41, pp. 103:1–103:26, Mar. 2023.

[9] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu, "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond," Apr. 2023.

[10] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018.

[11] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open Foundation and Fine-Tuned Chat Models," July 2023.

[12] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

[13] J. Liu, C. Liu, R. Lv, K. Zhou, and Y. Zhang, "Is chatgpt a good recommender? a preliminary study," *arXiv preprint arXiv:2304.10149*, 2023.

[14] K. Balog and F. Radlinski, "Measuring recommendation explanation quality: The conflicting goals of explanations," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 329–338, 2020.
[15] N. Tintarev and J. Masthoff, "Explaining recommendations: Design and evaluation," in *Recommender systems handbook*, pp. 353–382, Springer, 2015.
[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, 2022.
[17] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," *arXiv preprint arXiv:2203.11171*, 2022.
[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.
[19] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.