TABLE II. USER FEEDBACK EVALUATION RESULT

Metric	RAG	Fine-Tuned-GPT-4o	Combined
REL	3,35	3,51	3,54
ETU	2,97	3,16	3,32
EOU	2,95	3,3	3,35
INF	3,2	3,03	3,41
PRQ	2,93	2,99	3,22
TR	2,85	2,88	2,95
Overall Satisfaction	2,625	2,802	3,145

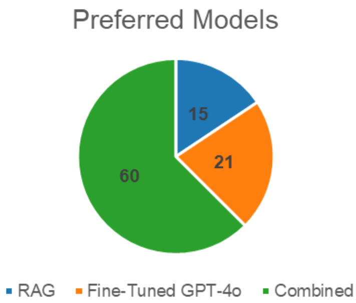

Fig. 5. Distribution of model preferences in user feedback evaluation

the Top-10 recommendation level, and 0.7393 at the Top-15 recommendation level, reflecting RAG's reliance on structured search.

The last metric in the quantitative evaluation is NDCG which assesses how well the recommendations are ranked in the Top-N recommendations by the CRS model. The evaluation results show that the Fine-Tuned GPT-4o model gets the best NDCG value at the Top-5 recommendation level with a score of 0.9951. This reflects the ability of Fine-Tuned GPT-4o in placing laptops according to their relevancy with the user's requirements. The Combined Model performed similarly with a score of 0.989 at the Top-5 level but had a small advantage at the Top-10 level, with a score of 0.9894, and Top-15 level, with a score of 0.9813. Meanwhile, the RAG model performed the worst at the NDCG metric with all scores on all recommendation levels being the lowest. This shows the limitation of the RAG model in maintaining the order of recommendation relevance on its list of laptop recommendations.

In addition to the Hit-Rate, Precision, and NDCG metrics, User Feedback can also provide a good overview of the practical usability and effectiveness of the model with real users. It can be seen in Fig. 5 that the Combined Model was selected 60 times by participants during the evaluation, while Fine-Tuned GPT-4o Model was selected 21 times and RAG Model 15 times. The preferences shown by users who have conducted the user feedback can also be supported by the results of Table II. Where the Combined Model gets the highest score on each evaluation criteria. The Fine-Tuned GPT-4o model follows with the second-best score across all criteria except informative, where it was surpassed by the RAG model. On the other hand, the RAG Model once again scored the lowest overall.

It can be seen from Fig. 4 that the Combined Model provides the best balance between Hit Rate, Precision, and NDCG metrics at each recommendation level, thus making it the best choice among the 3 models evaluated to provide relevant and accurate laptop recommendations. The Fine-Tuned GPT-4o Model followed closely to the Combined Model results but was overall inferior to it due to the limited data set, given that it did not use RAG's Retrieval Technique. And the RAG model is the worst model with the lowest scores in providing relevant recommendations and also has poor ranking. Table II and Fig. 5 furthermore shows the advantages of the Combined Model, where the Combined Model is the most preferred model by users with the highest user feedback scores on every evaluation criteria.

V. CONCLUSION AND FUTURE WORK

This research proposes the development of a Conversational Recommendation System (CRS) by utilizing Fine-Tuned GPT-4o with the retrieval technique of Retrieval-Augmented Generation to recommend laptops. To evaluate its effectiveness, we compared three CRS models: 1) RAG, 2) Fine-Tuned GPT-4o, and 3) Combined Model (Fine-Tuned GPT-4o + RAG's Retrieval Technique), using Hit Rate, Precision, and NDCG evaluation metrics at three recommendation levels (Top-5, Top-10, and Top-15). The evaluation results show that the Combined Model outperforms the other models across all metrics, making it the optimal choice for delivering accurate laptop recommendations. In addition to quantitative evaluation, the user feedback also offers significant understanding into the practical usability and effectiveness of the model with real-world users. The feedback shows that the Combined Model is the most preferred among participants with the highest scores on model selection and on each test criteria. This feedback further validates the results of the evaluation with Hit Rate, Precision and NDCG.

Although this study has laid a strong foundation by utilizing Fine-Tuned GPT-4o with the RAG search technique in CRS, several limitations remain. Despite its effectiveness, the search process is computationally demanding and may require optimization for faster response and further usability in real-time applications. This study also only concentrates entirely on the laptop domain, leaving its generalizability to other domains unexplored. In addition, the exploration of other LLMs, such as LLaMA or Alpaca, not just GPT-4o, may provide useful insights regarding their performance in CRS.

For future work or research tackling these limitations will be an important step. Using images or videos of the products that's being recommending could also improve the quality of the recommendation. Future research can also improve the validation of the model's evaluation by adding more User Feedback. These advancements could collectively refine the development of Conversational Recommendation Systems.

REFERENCES

[1] B. T. Imani and E. B. Setiawan, "Recommender system based on matrix factorization on twitter using random forest (case study: movies on netflix)," International Journal on Information and Communication Technology (IJoICT), vol. 8, no. 2, 2022, pp. 11–21, doi: 10.21108/ijoict.v8i2.655.

[2] Y. Sun and Y. Zhang, "Conversational recommender system," Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, pp. 235–244, 2018, doi: 10.1145/3209978.3210002.

[3] S. Dai et al., "Uncovering ChatGPT's capabilities in recommender systems," Proceedings of the 17th ACM Conference on Recommender Systems 2023, pp. 1126–1132, 2023, doi: 10.1145/3604915.3610646.

[4] S. Shahriar et al., "Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency," *Appl. Sci.*, vol. 14, no. 17, 2024, doi: 10.3390/app14177782.

[5] Y. Gao et al., "Retrieval-Augmented Generation for Large Language Models: A Survey," 2023, arXiv preprint arXiv:2312.10997. [Online].

Available: http://arxiv.org/abs/2312.10997

[6] M. S. Ayundhita, Z. K. A. Baizal, and Y. Sibaroni, “Ontology-based conversational recommender system for recommending laptop,” Journal of Physics: Conference Series, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012020.

[7] M. Cherian, E. Bobus, B. S. Jacob, A. S. Mandalika, A. M. Zacheria, “Empowering laptop selection with natural language processing chatbot and data-driven filtering assistance,” International Journal on Emerging Research Areas (IJERA), vol. 04, no. 01, pp. 364–371, 2024, doi: 10.5281/zenodo.12553277.

[8] H. Naveed et al., “A comprehensive overview of large language models,” 2023, arXiv preprint arXiv:2307.06435. [Online]. Available: http://arxiv.org/abs/2307.06435

[9] Z. Zhao et al., “Recommender systems in the era of large language models (LLMs),” 2023, arXiv preprint arXiv:2307.02046v6. [Online]. Available: http://arxiv.org/abs/2307.02046

[10] L. Ouyang et al., “Training language models to follow instructions with human feedback,” Advances in neural information processing systems (2022), vol. 35, pp. 27730–27744, 2022.

[11] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, “Improving language understanding by generative pre-training,” 2018. [Online]. Available: https://hayate-lab.com/wp-content/uploads/2023/05/43372bfa750340059ad87ac8e538c53b.pdf

[12] T. B. Brown et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, 2020, pp. 1877–1901.

[13] Z. K. A. Baizal, D. H. Widyantoro, and N. U. Maulidevi, “Factors influencing user’s adoption of conversational recommender system based on product functional requirements,” TELKOMNIKA (Telecommunication Computing Electronics and Control), vol. 14, no. 4, pp. 1575–1585, 2016, doi: 10.12928/TELKOMNIKA.v14i4.4234.

[14] A. Iovine, F. Narducci, and G. Semeraro, “Conversational recommender systems and natural language:: A study through the ConveRSE framework,” Decision Support Systems, vol. 131, Art. no. 113250, 2020, doi: 10.1016/j.dss.2020.113250.

[15] Z. K. A. Baizal, D. H. Widyantoro, and N. U. Maulidevi, “Design of knowledge for conversational recommender system based on product functional requirements,” 2016 international conference on data and software engineering (ICoDSE) , Bandung, Indonesia, 2016, pp. 1–6 doi: 10.1109/ICODSE.2016.7936151.

[16] D. Jannach, A. Manzoor, W. Cai, and L. Chen, “A survey on conversational recommender systems,” ACM Computing Surveys (CSUR), vol. 54, no. 5, 2021, pp. 1–36 doi: 10.1145/3453154.

[17] K. D. Spurlock, C. Acun, E. Saka, and O. Nasraoui, “ChatGPT for conversational recommendation: Refining recommendations by reprompting with feedback,” arXiv preprint arXiv:2401.03605, 2024. [Online]. Available: http://arxiv.org/abs/2401.03605

[18] Y. Liu et al., “Conversational recommender system and large language model are made for each other in e-commerce pre-sales dialogue,” Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 2023, pp. 9587–9605, 2023, doi: 10.18653/v1/2023.findings-emnlp.643.

[19] G. Zhang, “User-centric conversational recommendation: Adapting the need of user with large language models,” Proceedings of the 17th ACM Conference on Recommender Systems (RecSys ’23), Singapore, 2023, pp. 1349–1354, doi: 10.1145/3604915.3608885.

[20] W. Jiang et al., “PipeRAG: Fast retrieval-augmented generation via algorithm-System co-design,” arXiv preprint arXiv:2403.05676, 2024. [Online]. Available: http://arxiv.org/abs/2403.05676

[21] J. Jin, Y. Zhu, X. Yang, C. Zhang, and Z. Dou, “FlashRAG: A modular toolkit for efficient retrieval-augmented generation research,” arXiv preprint arXiv:2405.13576, 2024. [Online]. Available: http://arxiv.org/abs/2405.13576

[22] X. Wang et al., “Searching for best practices in retrieval-augmented generation,” Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 17716–17736.

[23] J. Kocoń et al., “ChatGPT: Jack of all trades, master of none,” Information Fusion, vol. 99, p. 101861, 2023, doi: 10.1016/j.inffus.2023.101861.

[24] Y. Hou et al., “Bridging language and items for retrieval and recommendation,” arXiv preprint arXiv:2403.03952, pp. 1–12, 2024. [Online]. Available: http://arxiv.org/abs/2403.03952

[25] S. Kemper et al., “Retrieval-augmented conversational recommendation with prompt-based semi-structured natural language state tracking,” Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 2786–2790. doi: 10.1145/3626772.3657670.

[26] J. Johnson, M. Douze, and H. Jegou, “Billion-scale similarity search with GPUs,” IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2019, doi: 10.1109/TBDATA.2019.2921572.

[27] L. V. Nguyen, Q. T. Vo, and T. H. Nguyen, “Adaptive KNN-based extended collaborative filtering recommendation services,” Big Data and Cognitive Computing, vol. 7, no. 2, p. 106, 2023, doi: 10.3390/bdcc7020106.

[28] G. Schröder, M. Thiele, and W. Lehner, “Setting goals and choosing metrics for recommender system evaluations,” UCERSTI2 Workshop at the 5th ACM Conference on Recommender Systems, vol. 23, 2011, p. 53.

[29] L. Gienapp, M. Fröbe, M. Hagen, and M. Potthast, “The impact of negative relevance judgments on NDCG,” Int. Conf. Inf. Knowl. Manag. Proc., 2020, pp. 2037–2040, doi: 10.1145/3340531.3412123.